Analytical, Big Data and Simulation Models of Railway Delays

Cerreto, Fabrizio

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Analytical, Big Data, and Simulation Models of Railway Delays

PhD thesis

Fabrizio Cerreto

June 2018

DTU Management Engineering
Department of Management Engineering

ANALYTICAL, BIG DATA AND SIMULATION MODELS OF RAILWAY DELAYS

PhD Dissertation

Fabrizio Cerreto
Department of Management Engineering
Technical University of Denmark

Supervisor:
Professor Otto Anker Nielsen
Department of Management Engineering
Technical University of Denmark

Co-supervisors:
Associate Professor Steven Harrod
Department of Management Engineering
Technical University of Denmark

Professor Bo Friis Nielsen
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Cerreto, F., 2018. Analytical, Big Data and Simulation Models of Railway Delays. PhD Dissertation. Department of Management Engineering, Technical University of Denmark, Kgs. Lyngby.

Cover photo: IDA Rail. https://universe.ida.dk/idarail/

PREFACE

This PhD thesis entitled *Analytical, Big Data, and simulation models of railway delays* is submitted to meet the requirements for obtaining a PhD degree at the Department of Management Engineering, Technical University of Denmark. The PhD project was supervised by Otto Anker Nielsen, Professor at DTU Management Engineering, and co-supervised by Steven Harrod, Associate Professor at DTU Management Engineering, and by Bo Friis Nielsen, Professor at DTU Compute. The thesis is paper-based and consists of the chapters listed in the tables of content, which include the papers listed below.

Paper I: Cerreto, Fabrizio. "Micro-Simulation Based Analysis of Railway Lines Robustness." In 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015), 164-1-164–13. Tokyo, Japan: International Association of Railway Operations Research, 2015.

Paper II: Cerreto, Fabrizio, Harrod, Steven, and Nielsen, Otto Anker. "A Closed Form Railway Line Delay Propagation Model." Re-submitted after the second round of review to *Transportation Research Part C: Emerging Technologies*, 2018.

Paper III: Cerreto, Fabrizio, Harrod, Steven, and Nielsen, Otto Anker. "Delay Estimation on a Railway-Line with Smart Use of Micro-Simulation." Edited by Gianluca Dell'Acqua and Fred Wegman. *Transport Infrastructure and Systems*, 2017, 867–74. https://doi.org/10.1201/9781315281896-112.

Paper IV: Cerreto, Fabrizio, Nielsen, Otto Anker, Harrod, Steven, and Nielsen, Bo Friis. "Causal Analysis of Railway Running Delays." In *World Congress on Railway Research (WCRR)*, 1–7. Milan, Italy: World Congress on Railway Research, 2016.

Paper V: Cerreto, Fabrizio, Nielsen, Bo Friis, Nielsen, Otto Anker, and Harrod, Steven. "Application of Data Clustering to Railway Delay Pattern Recognition." Journal of Advanced Transportation, 2018, 1–18. https://doi.org/10.1155/2018/6164534.

ACKNOWLEDGMENTS

A PhD is not only a challenge for the candidate but especially for all the people around. They take all the stress but do not receive a degree for this. I would like to at least acknowledge their contribution.

First and foremost, I would like to acknowledge the Innovation Fund Denmark and the industrial partners in the IPTOP research project, who made possible this PhD study. Special thanks go to Bernd Schittenhelm and Asger Purhus from Rail Net Denmark, and to Anders Vedsted Nørrelund and Michele Stawowy from DSB, who provided the data and deep insight into the rail operation and data analysis.

I am very grateful for my main supervisor Professor Otto Anker Nielsen and co-supervisors Associate Professor Steven Harrod and Professor Bo Friis Nielsen. Not only for the great help in the scientific content, but also for their guidance in my personal growth. Thanks to their daily support I learned to tackle difficulties and inconveniences without losing motivation.

Another major contribution to this PhD research project and to my personal development came from the stimulating environment that I found at the department of DTU Transport, currently DTU Management Engineering. In these years, I received a lot of comments and advice on my research from my colleagues and I was always able to establish an honest confrontation on any possible topic, within or outside transportation.

Thanks to all the people who became my family here in Denmark, who shared my experience and made me feel home. Thanks especially to Carola, Andrea, Enzo, and Giuseppe with whom I shared everyday stories for more than four years.

Thanks also to my actual family that I left in Rome. Thanks, because they always made me feel free. And thanks, because they were here when I needed them, despite I might have been a little absent.

Lastly, thanks to the special person that has shared with me most of my journey in the research, Luca. At any moment, I knew I could find support, motivation, and inspiration in him, and I am sure it just would have not been the same without. Thanks, because I always felt listened, and because he always made me get myself together whenever I was afraid of falling short. I look forward to being as supportive in your life challenges.

Fabrizio Cerreto, April 4, 2017

SUMMARY

Punctuality of railway networks depends on several factors, associated with the planning phase or the operation phase. In the planning phase, robust timetables are designed to withstand the variability in operation and to contain the generation of primary delays. Also, railway planners seek timetable stability to absorb primary delays reducing the propagation into secondary delays and return quickly to the unperturbed condition. Besides, primary delays that occur in the operations phase can be reduced by improving the industrial processes behind the railway services. Understanding how delays generate and propagate is central to the efficient design of robust timetables and corrective measures of service production processes.

The purpose of this study is the examination of the phenomena related to delays in railways, from both theoretical and empirical perspectives. The theoretical structure of delays is examined in analytical models. The effects of selected timetabling decisions are investigated in simulation models. Empirical studies on delay records from the realized operation are provided to identify recurrent patterns in the delay generation and recovery.

In the first section, the study evaluates commonly used indicators for timetable stability and robustness and compares their sensitivity to changes in traffic volume, heterogeneity, and the infrastructure layout. The comparison includes analytical measures based on the timetable structure and measures based on simulation of operation under known perturbations. On the one hand, ex-ante analytical measures focus typically on traffic heterogeneity and line exploitation, often considering individual characteristics of the timetable only separately. For instance, delay recovery is usually modeled through either running time supplements or headway buffers between trains. On the other hand, simulation of operation mimics the behavior of railway systems and provides a more detailed insight. Simulation tools allow different types of measurements, such as the individual train delays recorded at different timing points, which can be evaluated in different methods. The accuracy of simulation comes, though, at the price of higher demand for computational time and resources. In this section, aggregate delay as a function of primary delays is measured in a microsimulation environment, and it is described as a valid indicator of timetable reliability. However, the extensive calculation performed in microsimulation makes this method unsuitable for applications where the velocity of calculation counts. For instance, online applications for decision support tools need fast responses, in a few seconds, and heuristic optimization algorithms often require recursive calculations, so the overall response times dilate quickly. In this thesis, methods to reduce

the amount of simulation are also investigated, based on the same robustness measures under evaluation.

The first section of this thesis identifies a valid measure of timetable robustness in the aggregate line delay related to known incidents. One of the major obstacles to the application of this type of measure in real-time traffic management and optimization is its dependence on simulation, which is a time-consuming process. The following section presents alternative methods that combine analytical and simulation models to estimate the aggregate line delay as a function of primary delays with reduced resources requirement, paving the way to applications that require prompt responses. *In the second section*, an analytical model is presented to describe the delay propagation in a closed form function, allowing quick calculation of the reliability indicators identified in the previous section, including aggregate line delay. Analytical models are typically much faster than microsimulation and are therefore more suitable for optimization environments and online decision support tools. The mathematical model provides insight into the relationship between primary delays and the consequent total disturbance on railway lines. This relationship is described by a composite polynomial, which spans from first to third degree, depending on the magnitude of primary delay relative to the size of the study domain. Timetable design parameters can be adjusted in this model, and different settings can be quickly compared. The robustness given by different values of running time supplements, headway buffers, and punctuality threshold can be assessed. The model is initially formulated for homogeneous traffic on railway lines. It is later integrated with stochastic simulation to support heterogeneous traffic and to include the delay generation process. This process consists of three parts. The first part, the incident simulation, mimics events that block the railway, such a temporary track blockage, or signal failure, described by the distributions of initial time and duration. In the second part, the model generates primary delays combining the incident with the timetable structure. Lastly, the primary delay is propagated to the subsequent trains and the downstream stations. In the stochastic simulation model for heterogeneous traffic, the total delay is estimated as a consequence of an incident that affects an individual train service, and a weighted average is then used to derive the total delay function associated to the whole timetable. In addition to the aggregate line delay, the model provides the individual delays of every train recorded at each station and can be extended, therefore, to implement several metrics.

Both the analytical and simulation models presented in the previous sections rely on simplifying assumptions. One of the most influential assumptions, yet one of the most

frequent, is that trains always use all the slack available in the timetable to recover from delays, in the absence of further circulation conflicts. In reality, delay recovery is a stochastic process itself, and it is ruled by several factors, driving behavior, rolling stock performance, and passenger comfort among others. Furthermore, possible recovery depends on the allocation of timetable slack along the path. In the timetabling phase, railway planners typically allocate the slack according to general rules from practice. Investigation of recurrent patterns in delay development and recovery in railway operation can improve this process, giving the opportunity to tailor the slack according to specific characteristics of individual train services. The whole railway operation can also be improved identifying the factors that cause recurrent delays so that individual critical processed can be fixed, and specific delay mitigation measures can be designed. *In the third section*, this study lastly analyses empirical records from railway operation to extract information for modeling and to identify systematic delays that require specific countermeasures. Distributions of realized running times are studied to understand the real maximum performance of trains and the minimum feasible running time on a line section. The actual use of running time supplement to recover from delays highlights points of lack or excess of timetable slack. In this way, the real potential delay recovery available in the timetable can be determined to support robustness analyses of the timetable. Big data techniques are successively applied to empirical records to identify recurrent delay patterns to be associated with specific service characteristics, such as time factors and rolling stock performances. Timestamps from railway operation are arranged in delay profiles of individual service runs, which are then classified in clusters of services that develop their delay in similar ways. The method identifies locations where the delay changes recurrently in the same way, which may suggest changes in the schedules, or in the processes linked to the railway operation. The K-means clustering method finds application in very different fields, and it is generally appreciated for its simplicity and velocity. The resulting classes of delay profiles are eventually linked to the characteristics of individual trains, so that specific and focused corrective measures can be designed for the railway service production processes.

In summary, based on the knowledge developed in this study, it is possible to design robust timetables and to investigate the influence of selected parameters already in the planning phase. The study contributes the literature with an analytical delay propagation model, with the application of data analysis of the realized operation, and covers, besides, methods for appraisal of service reliability.

The total delay generated on a railway line as a function of primary delays is identified as the indicator that is most sensitive to variations in traffic volume and infrastructure improvements. Methods to estimate this measure without using microsimulation are proposed, making analyses quicker, and opening the possibilities to include such statistics in online applications and optimization models. Additionally, the empirical analyses presented permit the identification of recurrent delay patterns in railway operation, supporting the design of dedicated corrective measures of productive processes.

RESUMÉ (DANISH)

Rettidigheden af togtrafik på jernbanenet afhænger af flere faktorer som kan relateres til planlægningsfasen eller med driften. I planlægningsfasen bliver robuste køreplaner designet med fokus på at begrænse dannelsen af primære togforsinkelser samt på at absorbere dem for at hurtigt vende tilbage til normal drift ved at reducere opformering af sekundære forsinkelser af andre tog. Primære togforsinkelser i togtrafikken kan også reduceres ved at forbedre de bagvedliggende driftsprocesser. At forstå hvordan forsinkelser opstår og opformeres er centralt i forhold til at forbedre design af robuste køreplaner og korrigerende produktionsprocesser i jernbanetrafikken.

Dette studie undersøger de fænomener, der er forbundet med forsinkelser i jernbanedrift, både fra en teoretisk og empirisk vinkel. Den teoretiske struktur af forsinkelser bliver undersøgt grundigt ved hjælp af analytiske modeller. Effekterne af valgte beslutninger i køreplanerne vurderes med simuleringsmodeller. Endelig gennemføres empiriske studier af forsinkelsesårsager for at identificere tilbagevendende forsinkelser i togdriften.

I den første del af afhandlingen undersøges almindeligt anvendte indikatorer for køreplansstabilitet og robusthed for at forstå hvordan de er påvirket af trafikvolumen, heterogenitet af køreplaner og forskellige infrastrukturlayout. Den typiske fokus på trafikheterogenitet og linjeudnyttelse af ex-ante analytiske målinger sammenlignes med målinger baseret på driftssimulering. Driftssimulering efterligner driften af jernbanetrafikken og skaffer meget detaljeret indsigt herom, dog på bekostning af højere krav til beregningsmæssige ressourcer. Studiet anvender mikrosimulering af jernbanedriften og finder ud, at den aggregerede linjeforsinkelse, afviklingstid og gennemsnitlige togforsinkelse er egnede indikatorer for driftssikkerhed. Det bliver sammenlignet med standard indirekte målinger primært baseret på linjekapacitet og udnyttelse eller planlagte buffere mellem vognløb og deres fordeling. Den massive beregning, der kræves til mikrosimulering, gør dog metoden uegnet til online anvendelser til beslutningsstøtte såvel som til rekursive anvendelser i f.eks. de heuristiske optimeringsmodeller, der typisk bruges til køreplanoptimering. Afhandlingen undersøger derfor også metoder til at mindske mængden af simulation ved anvendelse af de samme målinger af robusthed under evaluering.

Som konklusion er mikrosimulering en særdeles detaljeret metode til at modellere jernbanedrift, men det er også en ressourcekrævende proces. For at mindske beregningsbehovet præsenteres derefter *i den anden del af afhandlingen* en analytisk

model til beregning af forsinkelsesopformering med en lukket formel, som tillader hurtig beregning af indikatorerne for driftssikkerhed. Analytiske modeller er typisk meget hurtigere end mikrosimulering og er derfor mere egnet til at indgå i optimeringsmodeller. Det er også tilfældet med den i studiet udviklede model. Den matematiske model skaber indsigt i forholdet mellem primære forsinkelser og den totale forsinkelse dannet på jernbanelinjer. Modellen beregner de individuelle forsinkelser af hver eneste tog ved hver station. Forholdet mellem primære forsinkelser og aggregerede linjeforsinkelser bliver vist i en sammensat polynom som spænder fra første til tredje grad ifølge forsinkelsesgenopretning. Køreplan parametre kan så justeres med denne model for at beregne effekten af forskellige værdier af køretidstilskud og togfølgetids buffere. Modellen udvikles først til homogen trafik på enkeltsegmenter af jernbaner og bliver derefter integreret med stokastisk simulation og udvidet til heterogen trafik samt til at omfatte forsinkelsesgenereringsprocessen. Fordelinger af afgangstid og varighed af begivenheder kombineres med køreplansstruktur for at modellere forsinkelsesgenereringen forårsaget af begivenheder, som for eksempel en midlertidig sporblokering eller en fejl ved signalerne. I den stokastiske model bliver den aggregerede forsinkelse beregnet som konsekvens af en begivenhed, der påvirker et specifik tog, og det vægtede gennemsnit benyttes derefter til at beregne den totale forsinkelsesfunktion forbundet til hele køreplanen.

Studiet analyserer *i den tredje del* empiriske data fra jernbanedrift for at uddrage oplysninger til parametrisering af modelleringen og for at kunne identificere systematiske forsinkelse, der kræver specifikke modforanstaltninger. Fordelinger af realiserede køretider studeres for at forstå togenes reelle maximale ydeevne og de korteste mulige køretider på en given strækning. På denne måde kan det faktiske slæk i køreplanen beregnes som støtte til robusthedsanalyser. Big-data teknikker anvendes til analyser af empiriske data for at identificere tilbagevendende forsinkelsesmønstre, som kan forbindes med specifikke serviceegenskaber, som f.eks. tidsfaktorer og ydeevne af det rullende materiel. Tidsstempler fra banedriften arrangeres i forsinkelsesprofiler af individuelle vognløb, hvilke så grupperes ved hjælp af cluster teknikker i grupper af vognløb med sammenlignelige forsinkelsesmønstre. K-means clustering anvendes i mange forskellige felter og værdsættes for metodens enkelhed og hurtighed. De resulterende klasser af forsinkelsesprofiler forbindes til individuelle vognløbs egenskaber, så specifikke og fokuserede korrigerende foranstaltninger kan designeres til jernbanedriften og køreplaner.

Alt i alt gør den udviklede viden det muligt at planlægge robuste køreplaner og at undersøge effekter af valgte parametre allerede i planlægningsfasen. Studiet bidrager til litteraturen med en analytisk model for forsinkelse udbredelse med anvendelse af dataanalyser for realiseret drift og dækker ydermere metoder for driftsikkerhedsvurdering.

Den totale forsinkelse genereret på en banelinje i forhold til primære forsinkelser vises at være den mest sensitive indikator for robusthed ved ændringer af trafikvolumen og infrastrukturforbedring. En analytisk metode udvikles til at beregne dette mål uden mikrosimulering. Denne metode gør analyser hurtigere og gør det muligt at inkludere sådanne mål i online anvendelser og optimeringsmodeller. Derudover tillader de præsenterede empiriske analyser identifikation af tilbagevendende forsinkelsesmønstre i jernbanedrift, hvilket støtter designet af specifikke korrigerende foranstaltninger af driftsmæssige processer.

TABLE OF CONTENTS

1 INTRODUCTION

The reliability of railway services is one of the most relevant factors that influence the attractiveness for passengers (Parbo et al., 2016). Beyond the expected magnitude of delays, the variability of travel times affects the passengers' preferences in the modal choice (Preston et al., 2009), and it can be measured, for instance, by the dispersion of delays. The increasing request for mobility is generating new challenges to the operators to keep adequate service quality while satisfying an enlarged demand. The relation between traffic volume and delays is, in fact, twofold. On the one side, an undersized service is often affected by primary delays at stations, which are generated by unplanned extensions of dwell times due to the large crowds (Huisman and Boucherie, 2001). On the other side, high traffic density entails a high degree of interactions between trains, which generates conflicts and secondary delays worsening the service quality (Gibson et al., 2002; Haith et al., 2014; Olsson and Haugland, 2004). Railway systems are inherently more constrained than other forms of transit, such as bus networks, and the infrastructure capacity constraints limit the number of transport services that can operate. Efficient use of the infrastructure, and of the transport system in general, is, therefore, especially significant in railways.

The combined theoretical and empirical knowledge about the interactions between the components of the railway system supports the development of more efficient plans and operations. A detailed comprehension of the generation, propagation, and recovery of railway delays facilitates an improved allocation of timetable slack and production resources. The safety and operational equipment integrated into the railway components collect data systematically and constitute an essential source of information for the planning, management, and revision of the processes. Empirical knowledge on delays can be deployed to improve the service punctuality, and therefore the attractiveness for passengers. The schedules may be designed more robust to the variations of daily operation, and corrective strategies may be implemented to improve the whole service production process and reduce thus the service time variability. Several research projects are focusing on methods to improve the transport service attractiveness, improving the

service reliability, increasing the transport supply, and improving the connections between different means of transportation (e.g., IPTOP[1], FOR2083[2], ONTIME[3], and Shift2Rail[4]).

This PhD project is part of the IPTOP research program at the Technical University of Denmark, which aims at improving the public transport by integrating the optimization processes among different operators, and across different means of transport.

1.1 Aim and main contribution

This PhD project focuses on rail operation, through data collection and analysis with mathematical and simulation models. The main purpose is to gain a better understanding of the formation, propagation, and recovery of delays in railways. Due to the research project constraints, the availability of material, and the opportunistic nature of data, the case studies of the distinct chapters focus on different Danish railway lines, or lines with similar characteristics from other countries, such as the Netherlands. The project contributes with insights from different perspectives, with the overall purpose of improving the service reliability of railways and their attractiveness for passengers. This includes theoretical models to describe how operational incidents develop into delays and how these delays propagate across services. Furthermore, empirical models presented in this dissertation describe the realized operation and identify systematic delays to be tackled with tailored corrective measures. The analyses presented in this manuscript are divided into three sections:

- Identification and comparison of measures of service reliability
- Analytical models of delay propagation in railways
- Data analysis of the realized operation

In the first section, different reliability measures are compared in terms of sensitivity to changes in the schedules and the infrastructure layout. Aggregate line delay as a function of primary delays results as an accurate estimate of the reliability of the timetable. This statistic is usually measured in simulation models, which are highly demanding for computational resources and difficult to integrate into optimization models.

[1] Research project "Integrated Public Transport Optimisation and Planning" funded by the Innovation Fund Denmark. http://www.iptop.transport.dtu.dk/

[2] Research project "Integrated Planning for Public Transportation" funded by the German Research Foundation. https://for2083.math.uni-goettingen.de/

[3] Research project "Optimal Networks for Train Integration Management Across Europe" funded by the Seventh Framework Programme of the European Union. http://www.ontime-project.eu/

[4] Research project "Shift2Rail" funded by Horizon 2020 from the European Commission. https://shift2rail.org

Therefore, an analytical model is presented in the second section to estimate the aggregate line delay in a faster analytical approach. The velocity of the model allows integration in environments that require a prompt response, such as recursive optimization models, or online decision support tools. Lastly, the simulation and analytical models are compared to real operation in the last section. The timetable design parameters for the analytical models, such as running time supplement and headway buffers, are, so, derived from the realized operation. The recorded timestamps show the share of available timetable slack that is deployed in reality to recover from delays. Furthermore, recurrent delay patterns are identified from the past operation and linked to the service characteristics so that tailored corrective measures can be designed.

The following paragraphs summarize the main characteristics, findings, and contributions of the individual sections.

1.1.1 Measures of service reliability

This section compares several measures of service reliability investigating their representation of the quality loss under perturbed operations. In particular, the focus of this study is on the reliability of the travel times in the daily operations, which are challenged by the natural variations of the industrial processes. Cancellations and major disruptions due to extreme events, such as snowstorms, are expectedly infrequent and are not considered in this survey. In fact, the aspect of the everyday-service reliability is often referred to by the terms Stability and Robustness. Stability points at the effectiveness of the timetable slack in absorbing the perturbations and taking the operations back to the undisturbed conditions. Robustness qualifies the goodness of the founding assumptions of a timetable, and their goodness to represent the real process-times distributions.

The stability and robustness of a timetable depend on its structure and can be improved by increasing the amount of slack scheduled or decreasing the degree of heterogeneity of the railway services. Indeed, the typical focus of the reliability measures available in the literature is on timetable heterogeneity and on the amount of slack scheduled. These aspects of the timetable structure can be summarized in descriptive analytical indices, which have the advantage of reduced computational cost in comparison to other methods. Other measures of reliability are based on the estimation of the cause-effect relationship and build often on the simulation of railway operations. In fact, the simulation of operation mimics the behavior of railway systems and provides a more detailed insight than analytical models. The price is, though, a considerably higher demand for computation and resources as compared to analytical models. Depending on the level

of detail of the simulation, this type of analysis often requires too long computation times to be suitable for fast applications. Examples are functions with recursive calculation, such as heuristic optimization models often used in timetabling, or the employment in online decision support systems that require fast responses. A graphical example of these two type of reliability measures is given in Figure 1.1-1. On the left side, the dispersion of headways between trains is used to estimate the degree of interaction, while on the right side, the individual train delays related to an incident are measured.

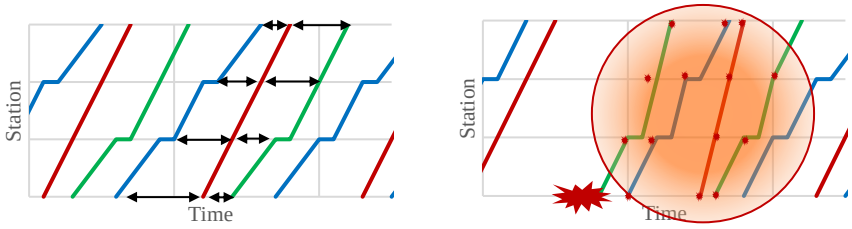

Figure 1.1-1: Comparison of descriptive measures of the timetable structure (left side) and measures of the cause-effect relationship under perturbed operation (right side).

Several reliability measures of transport operation are investigated in this section. The comparison between the measures focuses on the quality of their representation of the timetable's ability to withstand delays. In particular, the measures are confronted on their sensitivity to changes in the traffic volume and in the infrastructure layout. Both analytical and simulation-based measures are studied, including Total Amount of Running time Margin (TAoRM) (Salido et al., 2008), headway dispersion metrics (Carey, 1999), Sum of Shortest Headway Reciprocals (SSHR) and Sum of Arrival Headway Reciprocals (SAHR) (Vromans, 2005), Maximum Running time Difference (MRD) (Vromans et al., 2006), Weighted Average Distance (WAD) of running time supplements (Kroon et al., 2007), Heterogeneity measures (Haith et al., 2014; Landex and Jensen, 2013), Capacity consumption (UIC, 2004), Aggregate line delay (Barron et al., 2013), Settling time and average delay per train (Salido et al., 2012). The results of the comparison show that simulation-based measures, such as aggregate line delay, settling time and average delay per train, describe very well the consequences of disturbances in railway operation. The relationship between the magnitude of these initial disturbances, the primary delays, and the consequent total effect, assessed by the mentioned simulation-based measures, expresses the level of tolerance against perturbations of a timetable. These measures might be considered as an explicit enumeration of the effects of the incidents on individual trains. In fact, this characteristic makes simulation-based metrics a detailed and flexible tool to

describe how the service reliability would be affected by changes in the traffic volume and the infrastructure layout. However, the massive calculations required by microsimulation make it unsuitable for either online applications for decision support tools, or recursive applications like heuristic optimization algorithms. In this paper, methods to reduce the amount of simulation are also investigated, by sampling the cases to simulate. In heterogeneous timetables, the cause-effect relationship of disturbances depends, among others, on the specific train that receives the primary delays, which increases the number of cases to simulate proportionally to the number of different train services in the schedule. For example, in Figure 1.1-2, the case on the top stringline shows a local train delayed by 4 minutes, conflicting with an intercity train in the downstream section of the line. The case on the bottom stringline shows three different conflicts generated by the same primary delay assigned to a different train.

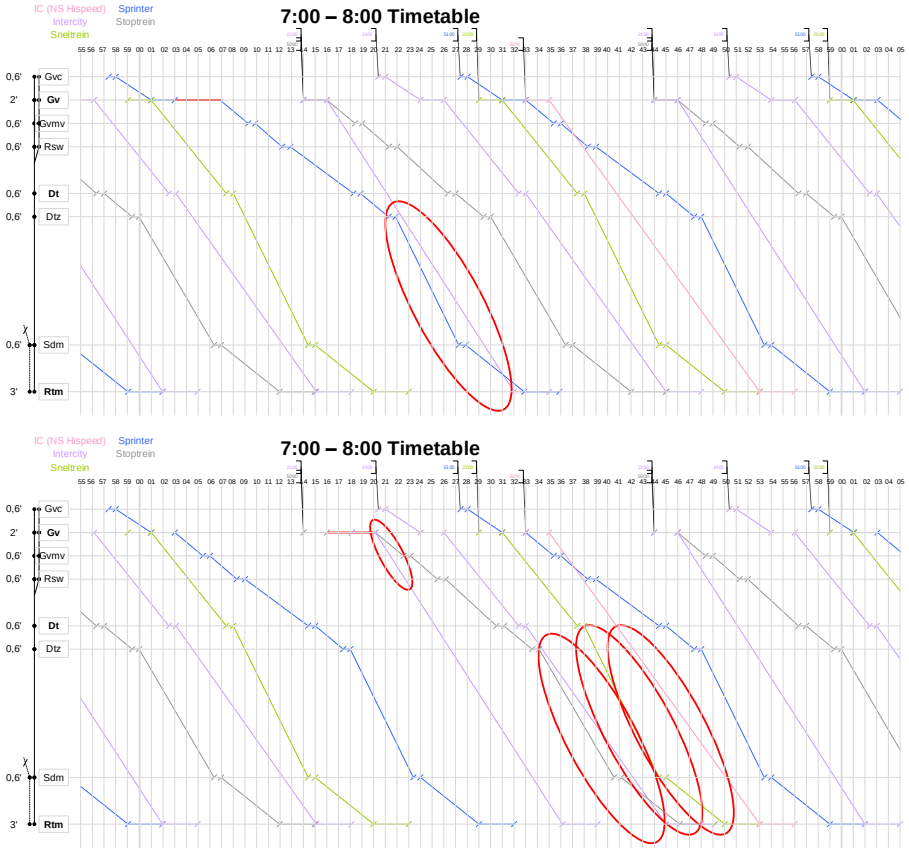

Figure 1.1-2: Comparison of the conflicts generated by assigning the same amount of primary delay to different trains in the same timetable.

The *skimming method* presented in this section approximates the overall effect of delays given to unspecified trains. In a sampled subset of the possible delay cases, the compares the cause-effect relationships linked to primary delays to the individual trains with the average effect linked to the timetable. The simulation of the entire pool of cases is then simulated assigning primary delays to a selection of trains. This method reduces the computational requirements to estimate the cause-effect relationship by introducing an approximation that should be assessed. The efficiency of the skimming method is directly related to the degree of heterogeneity of the timetable and yields the higher savings with the more heterogeneous schedules. Alternative methods to estimate the cause-effect relationship permit to reduce the computational requirement using mathematical approaches.

Given the suitability of cause-effect measures to describe the service reliability and the efficiency of analytical models, the following section introduces a new analytical delay propagation model on railway lines. This model mimics the results from microsimulation and returns the individual train delays in a railway system as a consequence of a primary delay. The aggregate line delay, the settling time, and the average train delay can be calculated in a closed form in much shorter time than microsimulation. It is, then, possible to integrate these measures in environments that require a prompt response.

1.1.2 An analytical model of delay propagation in railways

The previous section proposes explicit measures of the magnitude of perturbation as an indicator of the timetable reliability, analyzing the cause-effect relationship of disturbances on the schedules. Typical methods to measure the effects of disturbances include simulation of operation, which, especially at a high level of detail, entails massive calculation and long response times, resulting in limited applicability in high-paced environments. Alternative methods are presented in this section to estimate the consequences of disturbances in a faster way, through the analytical formulation of the relationship between individual train delays.

This section consists of two studies on the propagation of delays in railways in homogeneous or heterogeneous timetables. The overall aim is to develop faster methods to estimate the measures of reliability identified in the previous section. Analytical models are typically remarkably faster than microsimulation and are therefore more suitable to optimization environments and other contexts where short response time is relevant. The most significant contribution to the saving of time is given by simplifying assumptions on

the interactions between trains. In particular, a simplified recovery model is introduced for individual trains and across services, assuming pseudo-uniform running time supplements and headway buffer. The amount of details included in analytical models is, in fact, often reduced, and the faster calculation is also related to more approximated results (Mattsson, 2007; Meester and Muns, 2007).

A delay propagation model is presented in the first study to estimate the aggregate line delay, the settling time, and the average delay per train without simulation, as a result of a given initial delay (*A Closed Form Railway Line Delay Propagation Model*, re-submitted after second review to *Transportation Research Part C: Emerging Technologies*, 2017). The model provides the individual train delays recorded at single stations, and it can be extended to implement different types of aggregate metrics.

The model consists of two sections, which estimate the individual train delays as a function of a primary delay, and the aggregate line delay as a function of the individual train delays, respectively. The structure is shown in Figure 1.1-3.

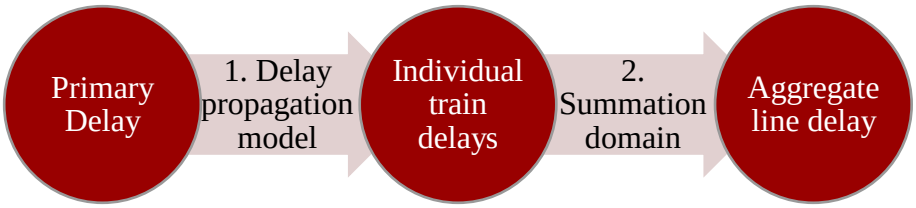

Figure 1.1-3: Scheme of the two sections of the delay propagation model.

In the delay propagation model, an initial delay is given to a train and propagated to the consecutive trains and the downstream stations if it exceeds the headway buffer or running time supplement, respectively. The delay of every train is calculated at every station as a combination of the disturbance provided by the previous train, or the residual delay from the previous station, as represented in Figure 1.1-4. The assumption of uniformly distributed slack makes the relationship between the consecutive delays of different trains at different location linear, which is practical for implementation in the second section of the model. Extending the linear relation between consecutive trains and stations, the delay recorded for every train at any station is thus determined directly from the primary delay and the scheduled slack.

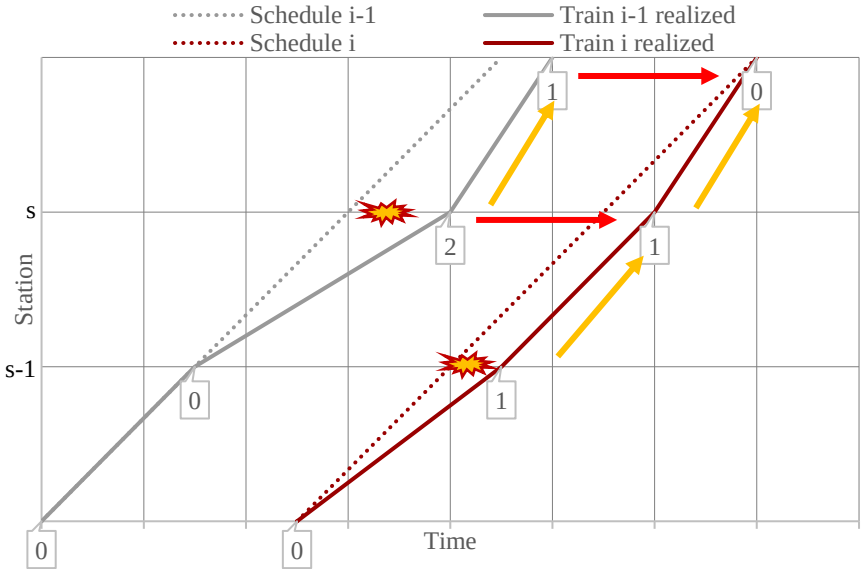

Figure 1.1-4: Scheme of the delay propagation model. The dotted lines are the scheduled trajectories, while the solid lines represent the realized operation. The first section of the analytical model aims at estimating the individual train delays at every station, which are reported in the callouts in the graph. The propagation through consecutive trains and stations is represented by the arrows in this figure. The stars represent primary delays.

The linear relationship between a primary delay and any individual train delay resulting from the first section defines a pseudo-triangular shape in the two-dimensional space of train services and stations, where the individual train delays are non-negative. The summation of the individual delays over this domain, named recovery region, defines the aggregate line delay. The recovery region is explored and divided into sub-regions, which boundaries define different types of relationship between primary and aggregate delay. The study region, represented in Figure 1.1-5, is the set of train services and stations included in the analysis.

The recovery region in the train-station domain is the region where trains run behind the schedule and it is defined as the set of services-stations with non-negative individual delays. The propagated individual train delays are independent of the study region, which makes the model flexible to different uses. In fact, the aggregate line delay is described as the summation of individual train delays, and different recovery conditions are defined by the boundaries of summation. The intersection between the study region and the recovery region results in the summation domain of the individual train delays. Such formulation makes the model flexible and applicable to different contexts, inclusive

of railway lines and networks, homogeneous and heterogeneous traffic, suburban and long-haul railway systems. In these cases, the service-station domain can be divided into homogeneous study sub-regions, where the timetable parameters and the traffic volume are constant. Multiple primary delays can be propagated recursively through the different homogeneous sub-regions of the system.

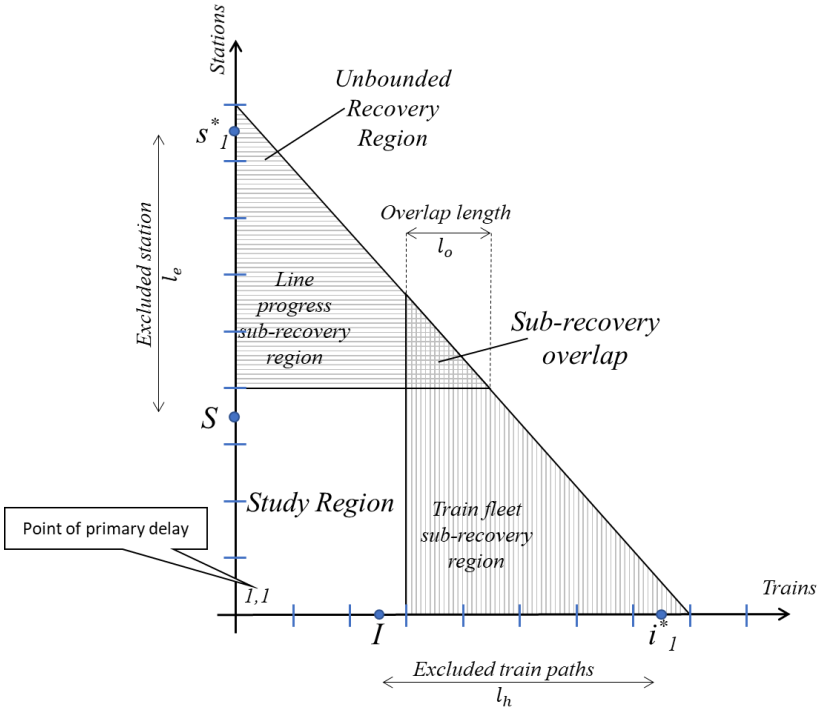

Figure 1.1-5: Study region and recovery region overlap in the service-station domain.

The most significant achievement of this study is that the cumulative delay can be calculated in a closed polynomial function of a primary delay. Such a differentiable formulation provides further information on the contribution of marginal increments of timetable slack in the damping of delay propagation. The differential calculus shows that the delay-damping effect of the timetable-slack decreases with its magnitude. Too large timetable slack does not improve sensibly the stability, while it still inflates the scheduled running times and headways between train, resulting in a reduction of the attractiveness for passengers.

In this first study, the analytical model is developed under the assumption of homogeneous or nearly homogeneous timetable, meaning constant stopping patterns and schedules across the services. Even though this is a common scheme in suburban railway

networks, extensive railway systems may result challenging to model when different types of service share the tracks. In the second paper of this section, the polynomial functional relationship between primary delays and aggregate line delay supports the extension of the model to heterogeneous timetables with a limited use of microsimulation.

In the second paper (*Delay Estimation on a Railway-Line with Smart Use of Micro-Simulation*, published in *Transport Infrastructure and Systems*, 2017), the analytical model is integrated with stochastic simulation to expand its applicability. This paper models the whole process of delay generation given by an incident, such as a temporary track blockage, or a signal failure. The model combines the distributions of initial time and duration of an incident with the timetable structure and returns the primary delay following an incident. Furthermore, the integrated stochastic model goes beyond the central assumption of homogeneous timetables from the purely analytical model presented in the previous paper. In heterogeneous timetables, the slack is not uniformly distributed across trains and stations, and the same primary delay given to different train services propagates differently and generates different perturbations. After the primary delay generation, the aggregate line delay is estimated in relation to the individual train services that receive the primary delay. The different functional relationships are then averaged to derive the total delay function associated with the whole timetable. In this first instance, microsimulation is introduced to measure the aggregate line delay corresponding to primary delays on individual train services, and the closed form function presented in the previous study is deployed to reduce the number of simulation runs. The two model extension of delay generation and weighted average are represented in the process flow in Figure 1.1-6.

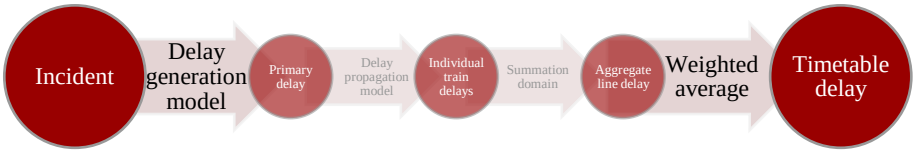

Figure 1.1-6: Extended aggregate line delay including the delay generation model and the overall timetable aggregate estimation. The newly introduced sections of the models are highlighted, as compared to Figure 1.1-3.

In the delay generation section, the primary delay is modeled as the result of the intersection of an incident and the timetable. A train receives a primary delay at a station if a blocking incident both starts before and ends after the scheduled departure. The delay generation model is represented in Figure 1.1-7.

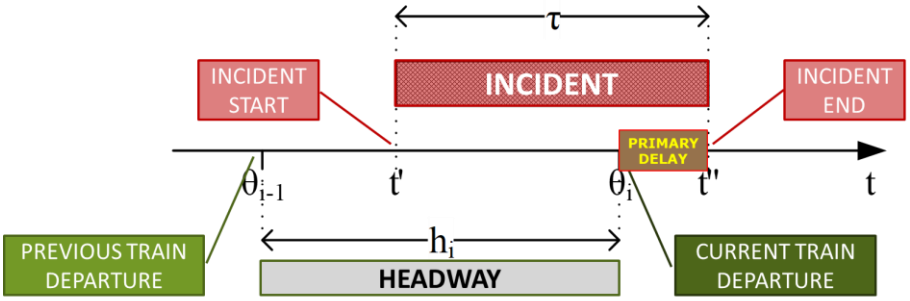

Figure 1.1-7: Delay generation model. The primary delay results from the intersection of an incident and the timetable, in the time dimension. In red, the incident main characteristics. In green, the timetable characteristics.

This model returns, in addition, the probability of a given incident to generate a primary delay on the individual scheduled trains. These probabilities constitute then the weights for the different functions of the aggregate line delay, corresponding to primary delays on different trains.

In future extensions of the model, the simulation might be abolished, and the parameters of the analytical model might be estimated from the timetable structure. The advantage of this approach is the possibility to account for heterogeneous timetables and to estimate the service reliability measures from distributions of incident times. Depending on the data recorded by railway operators, incident time distributions, may be of easier access than primary delay distributions. It is, in fact, somewhat challenging to isolate distributions of primary delays from recorded timestamps, while incident reports may result in a more straightforward collection of the disturbances durations.

The analytical model and its expansion in stochastic simulation presented in this section rely on the assumption that trains use all the slack available to recover from delays, if possible. In real operation, recovery is a stochastic process itself that varies across train drivers, rolling stock, and traffic management strategies. Especially under congested traffic, the interactions and interferences between trains put a limit on the individual train delay recovery. For instance, in the case of route conflicts, trains need to decelerate, stop and accelerate again according to the movement authority. These steps add time losses that consume part of the running time supplements and headway buffers and reduce therefore the possible delay recovery. Other factors that influence the recovery might be the promptness of the train drivers to adapt to the movement authority or the mechanical performance of the rolling stock, the visibility of signals, the weather conditions, and others. Furthermore, the models presented in this section rely on the availability of

timetable structural parameters, namely in the forms of running time supplements and headway buffers between trains. This type of information may be initially assumed or estimated in simulation models. However, the reliability of the plan depends also on the quality of the estimation of the slack available in reality. For example, the uncertainty in the calculation of the minimum feasible running times, or of the itinerary setup times, may lead to an underestimation of the running time supplement, or of the headway buffer, respectively. This uncertainty, and possible estimation, errors can be identified by analyzing data from the realized operation.

The next section proposes, therefore, methods to identify the best feasible performances in railway operation, so that the timetable slack available in reality can be estimated to feed the analytical and simulation models presented above. Furthermore, the theoretical insight into the relationship between timetable slack and the reliability of the service is confirmed in the analysis of real records from past operation. The historical data is also deployed to deepen the actual usage of running time supplements to recover from delays. These studies support the finding that too large running time supplement is unproductive. Not only too large slack does not contribute to damping the propagation of delays, but also it increases the running time variability and reduces, then, the reliability and attractiveness of the railway transport. The first study in the section highlights the stochastic nature of delay recovery. The second study on delay records finds recurring patterns in the variability of delay development and recovery.

1.1.3 Data analysis

The models presented in section 3 deal with delay recovery in a deterministic approach. Delayed trains are modeled to run at the maximum allowed speed and use all the timetable slack to reduce their lateness and stick to the schedule. Even though the implications of this assumptions are expectedly marginal at the aggregate level, the delay recovery process might differ across train services. Several factors can affect the ability of single trains to recovery from delays, such as the driving behavior, the dispatching strategies, the rolling stock performance, and other environmental factors. Furthermore, the inherent variability of the industrial processes and of the realization times, as opposed to deterministic schedules, makes the slack a stochastic variable itself. While the schedules are fixed, the minimum feasible process times may change according to several factors, taking the possible delay recovery on the stochastic level as well. An accurate estimation of the real process times (e.g. running times and itinerary setup times) reduces the uncertainty about the available slack and improves thus the robustness of the timetables.

Section 4 investigates historical data to extract information about the variability in the delay-recovery and seeks for recurrent patterns in the development of delays. The expected outcomes include a better representation of the delay propagation in the analytical and simulation models presented in section 3 and a better understanding of the delays in railways to tailor mitigation measures to improve the service reliability.

In the first paper (*Causal Analysis of Railway Running Delays*, published in the *Proceedings from the World Congress on Railway Research WCRR*, 2016), the distributions of realized running times on the busiest railway line in Denmark, Copenhagen – Roskilde, reveals the minimum feasible running times for different types of service. In the timetabling phase, the minimum running times are often estimated through either analytical formulation or microsimulation. Such estimation might result more or less accurate depending on the underlying assumptions and can be verified through historical data. The running time supplement included in the schedule might result under- or oversized in comparison to the actual rolling stock performances and the observed distributions of process times. The comparison between the revealed minimum running times and the scheduled running times returns the running time supplement available in reality. This is particularly relevant for the analytical delay propagation model presented in section 3, with the running time supplement being an input parameter, together with the headway buffer. Furthermore, the delays recorded for individual train journeys at sequential stations are compared. This study unveils, as a deduction, the existence of systematic delays related to dispatching strategies. For instance, trains traveling before schedule arrive at congested stations outside their designated time slot, which triggers dispatching decisions that very often lead to delays. The phenomenon is strongly correlated to excessive timetable slack, and over-recovery of. Typically, when trains reach the congested areas of the network ahead of the schedule, their designated station tracks are likely already occupied by other trains. This pattern leads, then, to late arrivals at congested stations, especially in case of reversing at terminus stations. This recurrent delay pattern is identified through a pairwise comparison of delay records at different stations. This type of analysis limits the amount of data that can be investigated and can only be applied to a few stations at one time. The development of further techniques opens the way to massive analyses to investigate longer train journey and to include several months of operation.

In the second paper (*Application of data clustering to railway delay pattern recognition*, published in the *Journal of Advanced Transportation*, 2018), big data

techniques are used to identify recurrent patterns in delay development and recovery. The purpose is to guide the railway planners towards more effective corrective measures to improve the service reliability and the attractiveness for passengers. The process is divided into the identification of systematic delays and the investigation of the influence of selected service characteristics. This operation has been traditionally operated by practitioners with consolidated knowledge on the specific lines under examination. The graphical representation of the delays was at the basis of the analysis, with the disadvantages linked to a laborious discrimination of the signals of systematic delays from the noise of random disturbances. A representation of the stacked delay profiles on the coastline between Helsingør and Copenhagen is provided, for example, in Figure 1.1-8. The main trend of delay increases towards the congested area of Copenhagen dominates the chart and hides other recurrent delay patterns.

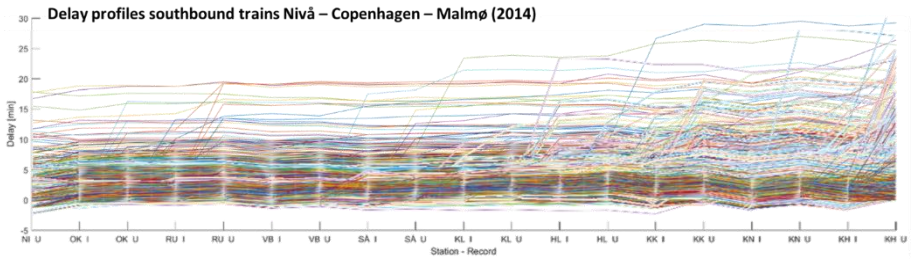

Figure 1.1-8: Delay recorded for individual train journeys towards Copenhagen. Stacked observations.

In this study, several observations of the same train service across different days are compared and partitioned in classes of similar elements using the *k-means* clustering algorithm. This algorithm is well known and has found applications in several fields of data analysis, but it is first introduced to delay profile analysis in this study. The algorithm performs a systematic classification which resembles the activity performed by expert analysts through the visual search for similarities in the observations. The advantage of such method is the freedom from biases and subjective interpretation of the observer, which makes it possible to examine large amounts of data. Cross data inference in the classes of observations reveals the factors that influence individual systematic delays. For instance, typical delay patterns are identified in conjunction with large passenger exchange at major stations in the peak hours, and other different patterns are only present in weekdays. Figure 1.1-9 represents the same dataset as Figure 1.1-8 clustered using the k-means algorithm. The different delay patterns are visible in the individual charts. In particular, the first two charts represent extensive increments of delays towards

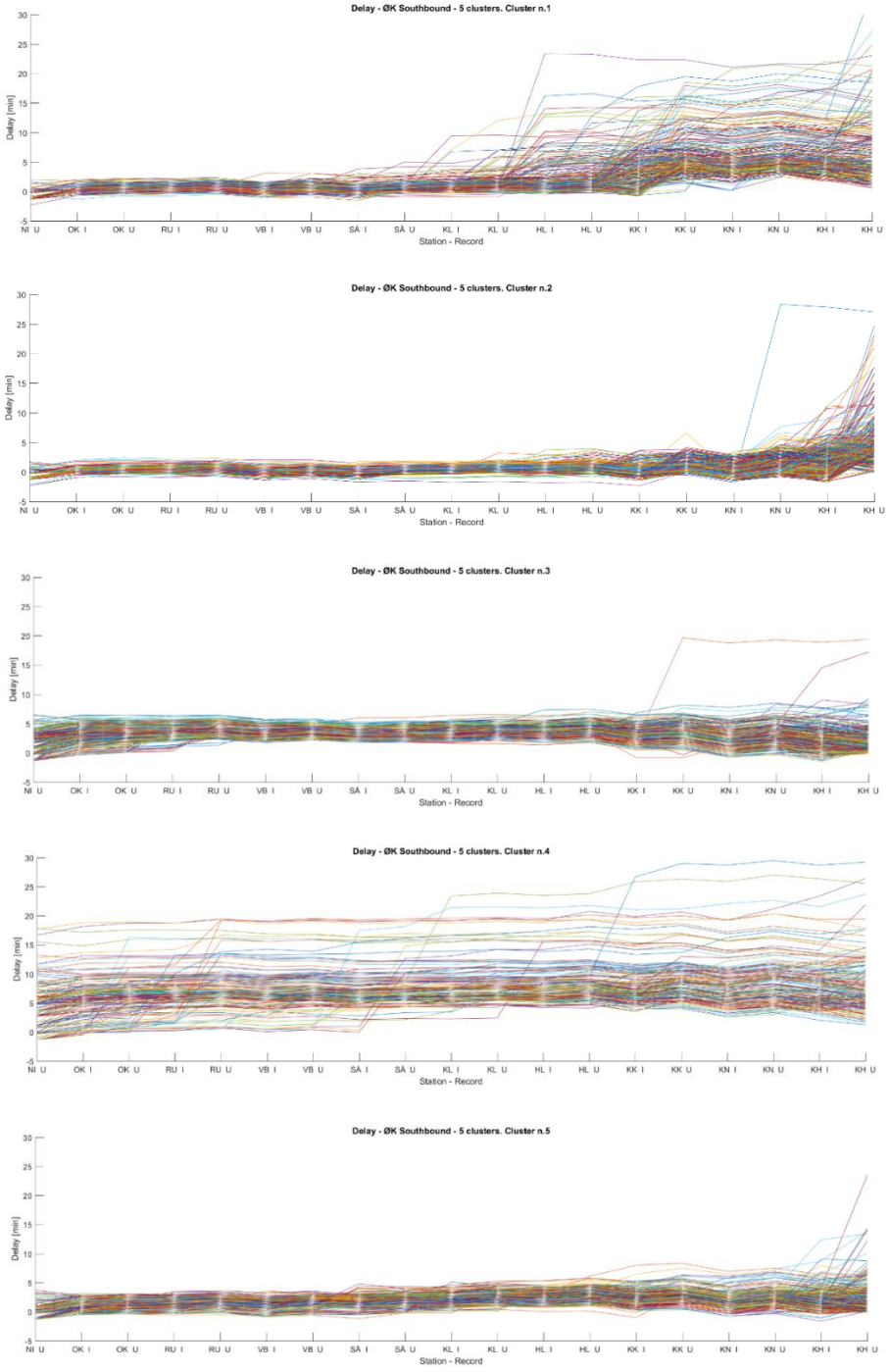

Figure 1.1-9: Delay recorded between Helsingør and Copenhagen, clustered according to recurrent delay patterns.

Copenhagen, recurrently taking place at specific stations on the line, which are also linked to different delay generating phenomena. The other graphs represent trains that are delayed through their entire journey, which slightly tend to either increase or decrease towards the destination. The same methodology is applied to the measures of delay change from the previous station. The inclusion of other sources of data might reveal other significant factors in the development of delays. Internal sources may be deployed, such as the delay reporting systems that include the causes of delays recorded by dispatchers. External sources such as registered weather conditions, or the passenger counts at the stations might be integrated into the factor analysis. The outcome of this research has a direct managerial impact on operations analysis, which can now point conditions that need specific corrective measure to mitigate delays.

The outcome of this section relates to the attractiveness of the railway service in two ways. The first study can be considered as a support of the analytical delay model presented in the previous section, whereas the second study relates directly with tactical decisions to correct those processes that cause the most significant extensions of running times and the systematic generation of delays.

1.2 Conclusions

This PhD study presents new insights into delays in railways, including delay generation, propagation, and recovery. The methodological contributions range from analytical models of delay propagation to techniques for data analysis of the realized operation, and include, moreover, a survey on measures for the assessment of the service reliability. The results of the research provide further knowledge on the composite-polynomial relationship between primary delays and aggregate line delays, on the variability of the delay recovery process, and on the relationship between excessive timetable slack and the generation of delays due to conflicts in operation. The methods cover theoretical approaches, simulation models, and analysis of real operation. Hence, this dissertation and the five associated papers contribute to the state-of-art within three main research areas of service reliability in railways: i) measures of robustness and their sensitivity to modifications in the system, ii) analytical models of delay propagation, and iii) analyses of realized operation to identify the actual slack in the schedules and recurrent delay patterns affecting the service reliability. The cases presented in the different sections of this manuscript range among Danish suburban, regional, and main railway line, and Dutch mixed traffic lines. The variety of case studies stems from the availability of data,

the opportunistic nature of data, and the relationship with other institutions and other research programs. The industrial and academic partners in the IPTOP research project made available different pieces of data following the interest of the institutions (e.g. DSB and RailNet Denmark are currently investigating methods to improve the unsatisfactory traffic reliability on the regional Coast railway line Helsingør-Copenhagen).

1.2.1 Measures of robustness

The findings of Paper I (chapter 2) highlight the suitability of simulation-based measures of robustness in assessing the possibility to withstand and absorb delays in a railway system. The study focuses on the sensitivity of such measures to modifications of the railway system, with the purpose of ranking different scenarios according to the improvement of service reliability. This is particularly relevant in the evaluation of changes in the infrastructure layout. These changes are typically highly onerous and require accurate analysis of the effectiveness in the improvement of the service. Among the simulation-based measures, the total delay generated on a railway line as a function of the primary delays is the most sensitive measure to variations in the traffic volume and the infrastructure. This aggregate measure is, therefore, the most suitable, among the investigated measures, for understanding the effects of variations in the service plan or in the operational settings. The results show a clear polynomial relationship between primary delays and aggregate line delay, whereas the settling time and the average delay per train follow a linear relationship to primary delays. This functional relationship is at the basis of the analytical model presented in the following section. As opposed to the measures of the cause-effect relationship, the compact and analytical metrics revealed considerably less sensitivity to changes in the service configuration. These statistics are commonly used in the industry, thanks to their simplicity and easiness to calculate, despite the approximate characteristics. Instances of these measures are the line exploitation and the capacity consumption, or the heterogeneity in the headways and running times. Indeed, most of the presented analytical measure only consider the average headway buffer, leaving out the running time supplement. The case study on a Dutch railway corridor shows that the same level of capacity consumption results from different timetables, where higher traffic is compensated by the homogenization of scheduled running times, which entails the reduction of the running time supplement for slower trains. In this way, the running time supplement is reduced, along with the possibility to recover from delays, but the measure of capacity consumption hides this weakness. The phenomenon is represented in Figure

1.2-1, where the timetable with 16 trains/h consumes less capacity than the timetable with 14 trains/h because of the shortened scheduled running times.

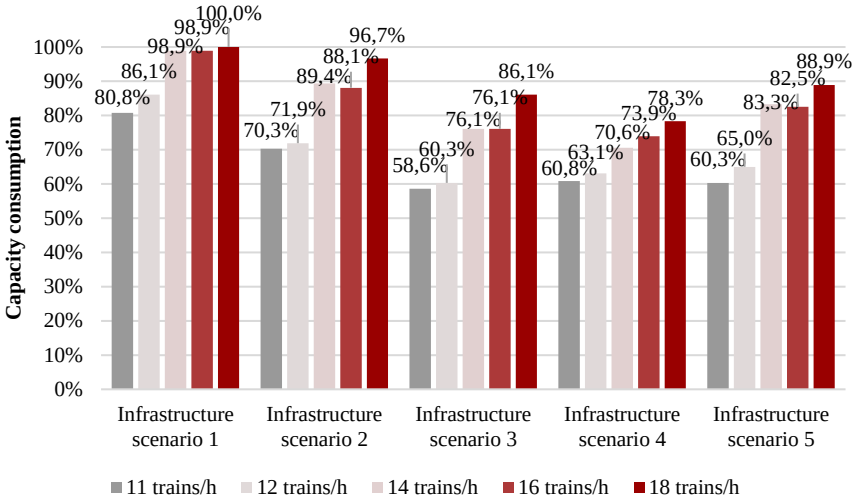

Figure 1.2-1: Comparison of capacity consumption for different infrastructure scenarios in relation to the traffic volume.

On the contrary, the drop of reliability linked to the decrease of running time margin is properly illustrated by the measure of aggregate line delay. Other measures based on the heterogeneity of either headways (normalized standard deviation and mean absolute deviation) or running times (maximum running time difference) are affected by increases of the traffic volume only indirectly, through the schedule modifications necessary to increase the capacity. The paradoxical result is that some dense timetables might appear more reliable than others that are considerably sparser if the latter are more heterogeneous.

Measures based on the estimation of the cause-effect relationship result thus more appropriate to measure the changes in reliability, and these are often based on the simulation of operation. One of the major obstacles to the introduction of simulation-based measures in online operations analysis is the long computation time. Micro-simulation models provide higher accuracy at the price of higher requirements for resources, both in the model design and in the analysis phase. In the paper, preliminary results are presented from methods to reduce the need for simulation, based on the selection of a sample of trains to receive primary delays.

Alternative methods to estimate reliability measures with reduced use of microsimulation are presented in the following section, based on the polynomial relationship primary delay vs. aggregate line delay identified in this section. Thanks to the

faster response from analytical models, the integration in optimization algorithms becomes possible, and the aspects connected to variations in reliability can be included already in the timetabling phase of railway planning. The analytical model for delay propagation finds its cornerstone in the functional relationship identified in the microsimulation of perturbed operation of this section.

1.2.2 Analytical models of delay propagation in railways

The results from microsimulation in the previous section highlight the polynomial relationship between primary delays and aggregate line delays, settling time and average delay per train. Thanks to the suitability of these measures to express the effects of changes in the railway system on the service reliability, and given the high resource demand of microsimulation models, this section presents alternative faster methods to estimate these measures analytically.

The purpose of chapter 3 is the introduction of a new analytical delay propagation model that allows a fast calculation of the measures identified in the previous chapter. The focus of both studies in this section is to remove or reduce the necessity of simulation in the calculation of total delay recorded on railway lines. Despite delay propagation models are known in the literature (Hasegawa et al., 1981; Landex, 2007; Pyrgiotis, 2012; Scheepmaker and Goverde, 2015), the delay recovery is often only partially modeled. In facts, the models presented in the literature often consider the timetable slack exclusively in the form of either running time supplement or headway buffer. In other cases, indirect measures of slack are introduced, such as the difference of speed between unperturbed and delayed operation, or the difference between actually scheduled and maximum theoretical train flow on the line. The model introduced in this chapter considers delay recovery through both scheduled running time supplement and headway buffer, and it provides a deeper insight into the functional relationship between primary delays and their effect on the overall operation.

The model is mainly developed in Paper II and offers a fast-analytic alternative to simulation, which makes the model suitable for online applications and recursive optimization environments. The closed form function provided in the paper is based on input design parameters that describe the timetable slack (running time supplement and headway buffers) and the aptitude of the operator in accepting small delays (delay tolerance threshold). The advantages of such formulation include the possibility to quickly evaluate the effects of different values of these control parameters. For example, the transport operator may desire to set the delay tolerance threshold to a value that reduces

the delays subject to penalties, according to the recorded delay distributions. This is particularly relevant in the tactical planning phase, where the infrastructure is defined and the possible changes concern mainly the timetable structure and the scheduled slack. Indeed, optimal values of timetable slack may be identified using this formulation.

The polynomial structure is consistent with the result from microsimulation of the first section. In fact, in Paper I, the relationship between aggregate line delay and primary delays is regressed to a second-degree polynomial, which corresponds to a partial recovery in the analytical model presented in this section. The simplicity of the base formulation, combined with a recursive application, allows yields the model flexibility and applicability to different delay scenarios. For instance, several simultaneous primary delays can be modeled on railway networks, including branching lines and sections with different traffic volume. Aggregate line delays, as well as settling time and individual delays for every train at individual stations, can be estimated quickly with good approximation. Moreover, other aggregate measures based on individual train delays may be derived from the linear delay propagation model. Theoretical insight on the propagation of delays in railways derives from the analytical model as well. In particular, the differential calculus highlights the reduction of effectiveness of the timetable slack in damping secondary delays. The main effect of too large headway buffers and running time supplements is, indeed, the extension of scheduled running times and the reduction of the service frequency, rather than the improvement of reliability, which ultimately reduces the attractiveness for passengers. Further analyses of the influence of the timetable parameters on the reliability strategic relevance of the choice of an appropriate tolerance threshold for delays by the operators. Passengers might not perceive, small delays, up to few minutes, and the operators can evaluate possible adjustments of the service contracts according to their own expectations of delays, and possibly save on the penalties for small perturbations to focus on the more sensible delays.

The analytical model presented in Paper II is demonstrated for homogeneous traffic, which is a common operational scheme, especially in suburban railways and metro system, or even on specialized high-speed lines. Nevertheless, the substantial heterogeneity of services on mainlines results more cumbersome to represent. The interactions between trains on the mainlines, in fact, vary along the route due to the relative differences in speed and stopping patterns, and the same primary delays given to different train services result in perturbations of different magnitude. In paper III, a stochastic

simulation model is presented to extend the analytical model, and include heterogeneous timetables.

The delay generation and propagation processes are modeled as an incident that generates a primary delay on a specific train, which propagates then to downstream stations and consequent trains. This approach extends the analytical model and yields the opportunity to estimate the aggregate line delay with heterogeneous timetables. The flexibility of the analytical model is maintained, and the applicability is extended to multiple real scenarios. Modeling primary delays may be difficult if their distributions are not available from historical data. In fact, the granularity of data may be too low to isolate primary and secondary delay distributions, and the actual records of specific incidents, such as signal failures, can be integrated into this model replacing the distributions of primary delays. The most significant advantage of such model is a considerable reduction in the simulation necessary to estimate aggregate measures of reliability in heterogeneous operation. The mixed simulation-analytical model deploys, in fact, the polynomial relationship identified elaborated in paper II to estimate the aggregate line delay as a function of primary delays. The stochastic simulation model included in this paper returns the weights of individual train services in the general aggregate line delay, which indicates the service reliability of the whole heterogeneous timetable.

At this stage, only a few simulations are required to estimate the polynomial relation specific of primary delays given to any individual train. Further development of the model may result in the accurate estimation of the delay recovery parameters from the timetable structure without the use of microsimulation and improve the applicability of the heterogeneous model in recursive algorithms as well as the homogeneous version.

1.2.3 Analyses of realized operation

The contribution of chapter 4 consists of new methods to extract information from historical data. This information provides insight into both the development of delays along the train journey and methods to identify and tackle systematic delays to improve the service reliability. The results of this section enforce the analytical model from the previous section with the identification of the actual running time supplement, and possibly headway buffer, which are the timetable parameters required in the estimation of the delay propagation. Furthermore, the findings on the stochasticity of the delay recovery support the conclusions from the analytical model. In particular, it is highlighted that too large timetable slack may even result as counter-productive due to the increased variability of the running times. Excessive timetable slack triggers possible delay development

phenomena, for instance when trains travel ahead of their schedule and result in congestion just before the major stations.

The first study in this section focuses on the identification of the minimum feasible running times on a railway section and on the actual use of running time margins recorded in past operation. The direct implication is the improvement of the timetable robustness and, thus, of the service reliability. The robustness, in fact, is defined in the literature as the quality the assumptions in a timetable, and the possibility to withstand variations in daily operation (Goverde and Hansen, 2013). The minimum running time is the basic component of the scheduled running times, which also include some running time margin to recover from possible delays. The minimum running times are often estimated by analytical models, or simulated through the vehicle dynamics. The estimation accuracy is strongly dependent on the quality of the assumptions in the formulation. For instance, these models should include the natural variability of the driving behavior or of the rolling stock performance linked to the environmental conditions. The analysis of historical data from operations provides the actual distributions of running times, which supports a more reliable estimation of the minimum feasible running times. The planners have, thus, the possibility to calibrate the analytical and simulation models against the real performances, and to schedule more reliable running times, for the timetable robustness benefit. Furthermore, the development of delays along the train journey is under focus in this paper, highlighting systematic delay patterns related to dispatching strategies. The recovery or increase of delays can be correlated to the effectiveness of the running time supplement so that the distribution of the slack can be tailored to the specific train runs. In fact, the results show when the running time margin is excessive, the variability of the running times increases reducing the reliability of the service. When the trains travel outside their designated time slot, these often arrive late at congested stations, even if they were traveling before schedule. The negative effect of early trains on reliability originates in the new route conflicts that these trains generate approaching stations when the previous train may still occupy their assigned track. Figure 1.2-2 shows this type of pattern recorded on the most congested railway line in Denmark, in the section from Roskilde to Copenhagen. The delays recorded at consecutive stations show that the majority of trains traveling early at the first station reach the last station behind the schedule. These results highlight the importance of well-designed and allocated running time supplements and headway buffers. In facts, not only too large supplements are not beneficial in delay recovery, as found from Paper II, but the higher probability of traveling before schedule

translates directly into higher probabilities of incurring in route conflicts at larger stations and generate more delays.

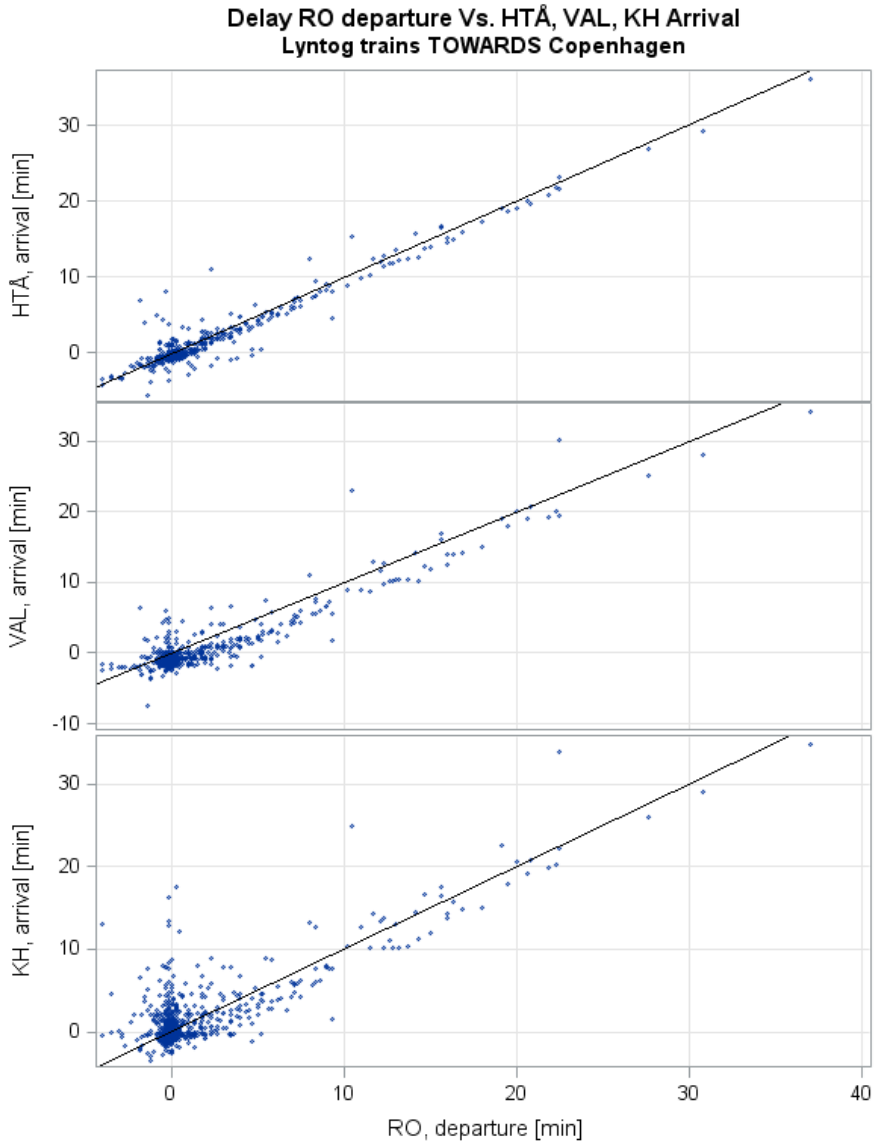

Figure 1.2-2: Recorded delays for long-distance services (Lyntog) towards Copenhagen central station. On the x-axes, the recorded delay passing the first station (Roskilde); points left to the 0-line describe trains traveling before the schedule at Roskilde. On the y-axes, the recorded delays at the downstream stations sorted top-down (Høje Tåstrup – HTÅ, Valby – VAL, Copenhagen central – KH). Among the trains recorded early at Roskilde, a considerable share is recorded late at Copenhagen central station.

This last result, in particular, raises the question about the existence of further patterns in the delay development and recovery. In real operation, though, several patterns, dependent on different factors, have effects on the same location, overlapping on the spatial dimension. For example, some patterns may be associated with long passenger exchange times at congested stations in the city centers, whereas other sections in the open lines may be more susceptible to the weather. This is, for example, the case of slippery rails in the rural areas due to the low temperatures and high humidity of the night. The aggregate statistical analyses deployed in this paper cannot distinguish between these different delay patterns from the recorded delays at the stations. Further methods are, therefore, explained and applied in the last paper in this manuscript, to investigate further recurrent delay patterns more systematically, and to relate these to specific service characteristics. Instead of comparing the delays recorded at pairs of stations, these new methods account for many more timing points at the same time and allow a more systematic and unbiased analysis of operation.

A Big-data analysis is applied in the second study of this section to identify recurrent patterns in delay development in train paths. This tool aims at supporting specifically the follow-up analysis of operation. In facts, the clustering algorithm applied seeks for internal structures in the dataset, meaning systematic repetitions of delays, which affect the service reliability and require mitigation measures. This operation has been traditionally left to the interpretation and experience of practitioners, who plotted the recorded delays of the individual train along the route and searched for similarities in the delay profiles. In this way, the analysis would easily be biased by artifacts in the plots, influencing the accuracy, especially with large samples. The accuracy of data-based algorithms is independent of the sample size, with the significant benefit that large datasets can be analyzed at once. For instance, data recorded in a whole year of operation might be inputted to the algorithm, highlighting systematic seasonal delays, which could not emerge by looking at single months. Thanks to the multiple sources of information integrated into railway systems, the quantity and quality of data are increasing. In future research, the cluster analysis may be combined with further sources of data to pinpoint the causes of the systematic delays identified. Data already available about the time of the day and the day of the week of individual trains highlighted the existence, in this case study, of systematic delays possibly related to the passenger flows. In facts, several train services were found systematically delayed at large stations in the pick hours according to the predominant passenger flows. Other patterns could potentially relate to systematic conflicts at junctions,

meaning that even small delays on trains from merging railway lines may have a considerable impact on the mainline punctuality. This classification tool facilitates the follow-up analysis and assists the operator in the design of tailored delay mitigation measures. The outcome is the expectedly improved effectiveness of the corrective measures and the improvement of the service reliability. In fact, an interview with the rail operator revealed that the planners implemented small changes in the headways on the line under study. The conflicts in the merging section were thus reduced and the punctuality improved beyond the operator's expectations. Further data from external sources, such as actual passenger flows, or recorded weather, may support the identification of additional causes of systematic delays, making it possible to design specific delay countermeasures.

Lastly, one of the main advantages of the method proposed is the transferability to other means of transportation, such as bus networks, or airlines. It is potentially applicable to any industrial process where the execution time can be compared against the schedule at given checkpoints.

1.3 Further research

While this dissertation contributes the literature with considerable progress regarding delays in railways, there is still ample room for improvement of the suggested analyses and algorithms and further new research to conduct within the topic.

The analytical model might be integrated with dispatching criteria to improve the accuracy in representing railway networks. The challenge will be keeping the simplicity of the model while introducing complex controls to mimic the prioritization strategies. Furthermore, in order to facilitate the integration in optimization algorithms, the analytical model may be implemented in automatized frameworks capable of converting a railway network structure into mathematical programming. This process would facilitate the recursive application of the model, making it a valid replacement of microsimulation in several contexts. With the same purpose, the mixed analytical and simulation model for heterogeneous networks can be further developed as well to eliminate the necessity for simulation in the estimation of aggregate line delay.

The research on the realized running times may be transported to the headways. In this way, the actual minimum headways between trains may be identified, together with the scheduled headway buffers. Minimum headways are in fact stochastic, similarly to the minimum running times. Even though the variability might be lower, there are still sources of uncertainty in the process time to guarantee a free path between two conflicting

movements. For instance, the variability in the feasible headways may be determined by the time to alter the position of the turnouts or the computational time in the interlocking to elaborate the signals, or by the length of the internal routes in the stations. However, possible changes in the order of trains in real operation require the analysis of the distribution of headways for all the possible permutations of train services, increasing the difficulty of the study. This condition is not valid in the study of running time supplement, as the order of stations is fixed for a train path.

Lastly, the study on recurrent patterns in railway delays may find application in other means of transportation. It is sufficient, in fact, a set of fixed checkpoints with a schedule and the related timestamps to generate delay profiles. This possibility applies to air traffic as well as bus and metro networks. Potentially, the application might be extended to several industrial processes that are not linked to transportation. At the same time, the implementation of additional sources of information would improve the understanding of the reasons for the systematic delays. Possible sources may include weather records, passenger counts, and onboard sensors for equipment monitoring.

The horizon of the big-data analysis might be expanded, including several train services at once. For example, time series of average delays recorded at the timing points across the day hours may be treated as individual observations. In this way, the resulting multidimensional delay profiles may be classified with the same method proposed in the last paper. These multidimensional delay profile would add a time-related dimension to the analysis, with the opportunity to investigate patterns in the delay propagation across services. The reliability of railway transport would be then improved tackling the factors that generate the most delay propagation.

1.4 Outline

The remainder of this thesis includes the five papers, divided into three thematic chapters. Hence, chapters 2 focuses on measures of robustness in railway transport, chapter 3 covers the two papers focusing on the analytical models to estimate the aggregate line delay, and chapter 4 includes the two papers on data analysis of realized operation.

References

Barron, A., Melo, P., Cohen, J., Anderson, R., 2013. Passenger-Focused Management Approach to Measurement of Train Delay Impacts, in: Transportation Research Board, 92nd Annual Meeting. Transportation Research Board of the National Academies, pp. 46–53. doi:10.3141/2351-06

Carey, M., 1999. Ex ante heuristic measures of schedule reliability. Transp. Res. Part B Methodol. 33, 473–494. doi:10.1016/S0191-2615(99)00002-8

Gibson, S., Cooper, G., Ball, B., 2002. Developments in transport policy: The evolution of capacity charges on the UK rail network. J. Transp. Econ. Policy 36, 341–354.

Goverde, R. M. P., & Hansen, I. A. (2013). Performance indicators for railway timetables. In 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings (pp. 301–306). IEEE. https://doi.org/10.1109/ICIRT.2013.6696312

Haith, J., Johnson, D., Nash, C., 2014. The case for space: the measurement of capacity utilisation, its relationship with reactionary delay and the calculation of the capacity charge for the British rail network. Transp. Plan. Technol. 37, 20–37. doi:10.1080/03081060.2013.844906

Hasegawa, Y., Konya, H., & Shinohara, S. (1981). Macro-Model on Propagation-Disappearance Process of Train Delays. Railway Technical Research Institute, Quarterly Reports, 22(2), 78–82. Retrieved from http://trid.trb.org/view.aspx?id=180725

Huisman, T., Boucherie, R.J., 2001. Running times on railway sections with heterogeneous train traffic. Transp. Res. Part B Methodol. 35, 271–292. doi:10.1016/S0191-2615(99)00051-X

Kroon, L.G., Dekker, R., Vromans, M., 2007. Cyclic railway timetabling: A stochastic optimization approach, in: Geraets, F., Kroon, L., Schoebel, A., Wagner, D., Zaroliagis, C.D. (Eds.), Lecture Notes in Computer Science. Springer, pp. 41–68. doi:10.1007/978-3-540-74247-0_2

Landex, A. (2007). Capacity Statement for Railways. Annual Transport Conference at Aalborg University, 1–19.

Landex, A., Jensen, L.W., 2013. Measures for track complexity and robustness of operation at stations. J. Rail Transp. Plan. Manag. 3, 22–35. doi:10.1016/j.jrtpm.2013.10.003

Mattsson, L.-G., 2007. Railway Capacity and Train Delay Relationships. Crit. Infrastruct. Adv. Spat. Sci. doi:10.1007/978-3-540-68056-7_7

Meester, L.E., Muns, S., 2007. Stochastic delay propagation in railway networks and phase-type distributions. Transp. Res. Part B Methodol. 41, 218–230. doi:10.1016/j.trb.2006.02.007

Olsson, N.O.E., Haugland, H., 2004. Influencing factors on train punctuality—results from some Norwegian studies. Transp. Policy 11, 387–397. doi:10.1016/j.tranpol.2004.07.001

Parbo, J., Nielsen, O.A., Prato, C.G., 2016. Passenger Perspectives in Railway Timetabling: A Literature Review. Transp. Rev. 36, 500–526. doi:10.1080/01441647.2015.1113574

Preston, J., Wall, G., Batley, R., Ibáñez, J.N., Shires, J., 2009. Impact of Delays on Passenger Train Services. Transp. Res. Rec. J. Transp. Res. Board 2117, 14–23. doi:10.3141/2117-03

Pyrgiotis, N. (2012). A Stochastic and Dynamic Model of Delay Propagation Within an Airport Network For Policy Analysis. Massachusetts Institute of Technology. Retrieved from https://dspace.mit.edu/handle/1721.1/71452#files-area

Salido, M.A., Barber, F., Ingolotti, L., 2012. Robustness for a single railway line: Analytical and simulation methods. Expert Syst. Appl. 39, 13305–13327. doi:10.1016/j.eswa.2012.05.071

Salido, M.A., Barber, F., Ingolotti, L., 2008. Robustness in railway transportation scheduling, in: Proceedings of the World Congress on Intelligent Control and Automation (WCICA). IEEE, Chongqing, China, pp. 2833–2837. doi:10.1109/WCICA.2008.4594481

Scheepmaker, G. M., & Goverde, R. M. P. (2015). The interplay between energy-efficient train control and scheduled running time supplements. Journal of Rail Transport Planning & Management, 5(4), 225–239. https://doi.org/10.1016/j.jrtpm.2015.10.003

UIC, 2004. Leaflet 406 - Capacity.

Vromans, M., Dekker, R., Kroon, L.G., 2006. Reliability and heterogeneity of railway services. Eur. J. Oper. Res. 172, 647–665. doi:10.1016/j.ejor.2004.10.010

Vromans, M.J.C.M., 2005. Reliability of Railway Systems. Netherlands TRAIL Research School.

2 MEASURES OF THE RELIABILITY OF RAILWAY SERVICES

2.1 Paper I: Micro-Simulation Based Analysis of Railway Lines Robustness

Cerreto, Fabrizio. "Micro-Simulation Based Analysis of Railway Lines Robustness." In 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015), 164-1-164–13. Tokyo, Japan: International Association of Railway Operations Research, 2015.

The paper presented below is the result of a major revision after publication in the conference proceeding from RailTokyo2015.

Abstract

Railway Undertakings and Railway Infrastructure Managers have a variety of parameters to measure the robustness of timetables: this paper examines empirical data collected from Nederlandse Spoorwegen on the heavily occupied railway line between The Hague and Rotterdam in The Netherlands. The results show that the robustness indicator examined are affected by the traffic volume and other timetable characteristics in different ways.

Analytical and micro-simulation-based measures of timetable robustness are applied to different railway infrastructure scenarios and compared to common measures such as the capacity consumption, and the share of trains delayed in case of disturbance. The relationship between simulation-based measures and the primary delays is estimated through regression analysis or differential calculus. The sensitivity of these measures to increases of traffic volume is consequently investigated through an amplification factor as a function of the train frequency. A skimming method is used for the sampling of simulation scenarios to reduce the computational time. The benefits of modifications to the track infrastructure, the timetable, and the signaling system, in terms of consecutive delays reduction, are estimated by giving a range of primary delays to a selection of trains.

The research highlights the need for a step further than currently planned in the infrastructure development to improve the line's robustness.[5]

The findings are significant for the relationship between IMs and RUs, as the same infrastructure or planning/scheduling improvements could be measured in a different way from each other contractor, with an economic impact on the infrastructure use agreements.

KEYWORDS: *Stability, Robustness, Microsimulation, Timetable, Railway infrastructure, Delays*

[5] On February 8 2018, ProRail, the Dutch railway infrastructure manager, announced plans to furtherly upgrade the line to increase the train frequency. The decision is in agreement, in fact, with the conclusions of this paper from 2015. See: https://www.globalrailnews.com/2018/02/05/e300m-upgrade-for-the-hague-rotterdam-rail-route/ and https://www.railwaygazette.com/news/infrastructure/single-view/view/den-haag-rotterdam-upgrade-to-support-a-train-every-5-min.html

2.1.1 Introduction

Investments in railways usually require massive resources from both IMs (Infrastructure Managers) and RUs (Railway Undertakings): alignment modification and signaling system upgrades on one hand, and rolling stock renovation on the other, should be carefully designed and examined. Therefore, every modification needs benchmarking and measures of the actual results. One of the most relevant aspects of the improvement of operation quality is the timetable robustness, especially on densely occupied networks. Several measures of robustness exist, with a focus on different aspects of the disturbances in the daily operation or of the planning tools adopted to mitigate delay propagation. IMs and RUs can choose on a variety of indices to assess robustness, which are influenced by increases of traffic volume and consider primary delays in different ways: this paper analyses the relation between selected measures of robustness, the traffic volume, and the primary delays. The analysis is based on micro-simulation, thus resulting in a resource-intensive process. A method is thus proposed, in addition, to reduce the computational load with a reasonable approximation of the simulated results.

A survey on the different definitions, measures of robustness and the related methods is presented in section 2.1.2. The comparison method is described in section 2.1.3, where the microsimulation tool, the procedures to examine the traffic volume influence and to reduce the computational load are explained. In section 2.1.4, the method is applied to the railway corridor between The Hague and Rotterdam, in The Netherlands. General conclusion for the method proposed and the possible further research are given in section 2.1.5.

2.1.2 Survey on the robustness of timetables

2.1.2.1 Robustness definitions

Regional and suburban railway networks are often characterized by high traffic density and heterogeneity of services and are thus sensitive to disturbances. High service frequency implies short headways and limited buffer times between scheduled services, with considerable influence on delay propagation. Traffic density and occurrence of disturbances in railway operation are often in positive correlation (Wiklund, 2002), and the extent of disruptions is also strongly affected by the traffic volume (Gibson et al., 2002; Haith et al., 2014; Jensen, 2015).

Several definitions of robustness in railway operations exist, with a focus on different aspects of reliability, and all related to the propagation of delays. Some of the literature refers to the general ability to absorb delays through the timetable slack (Andersson et al.,

2011). This feature is also referred to as Internal Robustness (Hofman et al., 2006) or stability (Goverde and Hansen, 2013). Andersson et al. (2013b) investigate the robustness of timetables analyzing the timetable slack in critical points, paying particular attention to the flexibility of operation in terms of feasible dispatching strategies. Peterson (2012) compared different strategies for timetable slack allocation in a micro-simulation environment and identified the most robust schedule as the one resulting in the highest punctuality.

In other cases, robustness is described as the ability of a timetable to withstand the variations in daily operation, which is given by accurate estimation of process times and primary delay distributions (Goverde and Hansen, 2013).

Robustness is directly connected to the interplays between the timetable, the infrastructure, and the rolling stock characteristics. Strategies to improve the timetable robustness include the increment or the intelligent allocation of timetable slack (Peterson, 2012; Schittenhelm, 2011; Solinen et al., 2017), the reduction of heterogeneity of services in general (Salido et al., 2008), the homogenization of headways (Vromans et al., 2006), and the containment of differences in scheduled running times (Huisman and Boucherie, 2001).

2.1.2.2 Robustness measures

The differences of the business targets between Railway Undertakings and Infrastructure Managers drive different strategies to improve the service reliability. RUs tend to favor the increase of buffer times between trains, as the running time supplements increase the scheduled running times and drop the service appeal to the passengers. On the other hand, IMs make profits from the sale of train paths, which availability is reduced by the extension of buffer times. Many parameters are available to measure robustness, according to the purpose of the performance analysis, and, to our knowledge, there is no literature on the negotiation on robustness performance: different KPIs better suit the point of view of either RUs or IMs. In addition, there is an increasing interest in maximizing the use of railway capacity, with benefit to both the RUs, which can operate more trains and increase ticket revenue, and the IMs, with a direct increase of income given by additional slots available. The effects of additional train paths on service reliability have been studied by Haith et al. (2014) in a framework to define a congestion charge for RUs applying for additional slots on congested lines. The international development of the railway markets is shining a light on the need to integrate the railway systems across different countries, legislations and technological infrastructures. Therefore, the need for a shared method to

evaluate the performance in robustness is rising: interoperability on one side, and the diversity of railway systems on the other side, require an approach adaptable to different contexts.

Different techniques to evaluate railway robustness have been proposed in the last years, some of them based on analytical approaches, others based on empirical models and what-if analyses. The advantage of analytical measures is their simplicity and their quick calculation, with limited need for information, which makes these measures best fit for initial evaluations in the planning process where detailed information on the individual services is not available (Meester and Muns, 2007). The simplicity comes, though, at the cost of accuracy. Indeed, these measures often include considerable approximation and may not be suitable for accurate analyses. Microsimulation models, on the contrary, provide a high level of detail, but are highly resource-intensive and require much time to both build and operate (Carey, 1999; Carey and Kwieciński, 1994; Parbo et al., 2014).

Analytical measures of robustness

Analytical estimation of robustness is often based on the evaluation of heterogeneity among the scheduled services. Carey (1999) proposed a number of reliability metrics at a railway station, divided into metrics that require knowledge of primary delay distribution functions, and metrics based only on timetable measures. In case these distributions are unknown, Carey used two metrics based on the dispersion of headways in the timetable.

$$1 - \frac{s.d.\ of\ headways}{\frac{H_T}{n}} \tag{1}$$

$$1 - \frac{m.a.d.\ of\ headways}{2\frac{H_T}{n}} \tag{2}$$

with H_T being the total headway available and n the number of headways.

Vromans et al. (2006) proposed further measures of heterogeneity in the scheduled headways, which also consider the differences in the scheduled running times on line sections.

$$SSHR = \sum_{i=1}^{n} \frac{1}{\bar{h_i}} \tag{3}$$

$$SAHR = \sum_{i=1}^{n} \frac{1}{h_i^A} \tag{4}$$

where h_i^- is the shortest headway between trains i and $i+1$ on a line section, and h_i^A is the arrival headway at the end of the section between the same trains.

Based on the metrics proposed by Vromans et al., Haith et al. (2014) proposed further measures to account for the actual headway buffer included in the timetable:

$$HET_S = \frac{SSHR}{\frac{1}{HW} \times g} \times 100 \tag{5}$$

$$HET_A = \frac{SAHR}{\frac{1}{HW} \times g} \times 100 \tag{6}$$

The slack scheduled in the headways is accounted in these formulations by means of HW, which is the minimum feasible headway given by the distancing system, while g is the number of headways considered. The measure is thus relative and spans in the range 0%-100%.

Landex and Jensen (2013) also proposed a set of metrics inspired by Vromans et al., which are normalized and eliminate the dependency on the traffic volume, and focus on the departures from, or the arrivals at a single station.

$$Het_A = 1 - \frac{\sum \min\left(\frac{h_{t,i}^A}{h_{t,i+1}^A}; \frac{h_{t,i+1}^A}{h_{t,i}^A}\right)}{h_{N-1}} \tag{7}$$

$$Het_D = 1 - \frac{\sum \min\left(\frac{h_{t,i}^D}{h_{t,i+1}^D}; \frac{h_{t,i+1}^D}{h_{t,i}^D}\right)}{h_{N-1}} \tag{8}$$

Other measures of schedule heterogeneity were presented by Andersson et al. (2013a), with focus on the scheduled running time differences, instead. The presented measure is named Maximum Runtime Difference and is given by

$$MRD = \max(R_i) - \min(R_i) \,|\, i \in I \tag{9}$$

where $i \in I$ are the individual trains in the schedule, and R_i are the related scheduled running times.

Andersson et al. (2013a) compared several of the robustness measures listed above and proposed a new measure that considered simultaneously two different forms of timetable slack, which are the headway buffers and the running time supplements at critical points, an. The proposed measure RCP (Robustness at Critical Points) expresses the flexibility of operation available to the dispatcher to manage the rail traffic in perturbed operation and it is shown to be a valuable measure of reliability. However, this measure focuses on specific critical points in the space-time domain, identified on the basis of potential conflicts between trains. The identification of the defined critical points follows a structured procedure, but it might miss interactions between specific trains that are not identified as critical.

A different stream of research on measures of robustness focused particularly on the timetable slack incorporated in the schedules. Salido et al. (2008), for instance, computed the total amount of running time supplement available in a schedule.

Simulation-based measures of robustness

A number of robustness measures have been proposed based on simulation of the timetable. The simulation was used to assess the available slack in the timetable, with special regards to the headways, or combined with distributions of primary delays to assess the ability of a timetable to recover from the disturbed operation.

The capacity consumption was promoted by the International Union of Railways (UIC, 2004). The measure expresses the line exploitation and measures the average headway buffer in the timetable and is associated with recommended maximum values to contain delay propagation. The line capacity consumed by a timetable is calculated as the ratio of the minimum time occupied by a compressed version of the same timetable divided by the total scheduled time,

$$\eta = \frac{t_e}{t_p}, \tag{10}$$

where η is the capacity consumption, t_e is the compressed time of line exploitation, and t_p is the scheduled period. While the UIC capacity consumption is measured at the level of blocking sections, possibly using micro-simulation, the CUI (Capacity Utilization Index) is a similar measure, at a macroscopic level (Gibson et al., 2002; Haith et al., 2014; Mattsson, 2007).

The relationship between primary delays and the related overall disturbance effect on railway operation was at the base of several measures of robustness with micro-simulation tools. Salido et al. (2008) simulated incidents on a railway line and assessed

robustness in terms of the number of trains delayed, average delay per train, and settling time, to evaluate different solutions of their rescheduling model. The settling time was defined as the time necessary to absorb a given perturbation so that all trains are recorded on time. This measure was also identified by Goverde and Hansen (2013) to assess the ability of a timetable to absorb perturbations.

The total, aggregate, or cumulative line delay recorded as the effect of known disturbances is also a common measure of the ability of the railway system to withstand delays. This type of measure is also often used as an ex-post performance measure and is pointed as one of the most representative of the realized service reliability (Barron et al., 2013). Harker and Hong (1994) assessed the quality of a dispatching management algorithm measuring the total deviation recorded on a line. Ginkel and Schobel (2007) evaluated the quality of their bi-criteria delay management algorithm by means of the total delay generated by decisions of keeping or ignoring the service connections in case of disturbances. Corman et al. (2014) compared the robustness of different simulated scenarios in perturbed rail operation measuring, among others, the average total delay generated by stochastic primary delays. Solinen et al. (2017) benchmarked a selection of robustness measures against different forms of aggregate delays recorded in a micro-simulation environment.

The methods listed in this survey measure robustness of singular solutions, but their relationship to the train frequency is yet to be understood: in this paper, we propose an evaluation of the link between selected robustness indicators and the increment of train frequency. The results show that some indicators are more affected by traffic volume increases than others: general conclusions about the disruptions propagation and fade are shown in the last section.

2.1.3 Methods

This study focuses on robustness appraisal in relation to the traffic volume. The magnitude of given initial disturbances is put in relation with the overall effect on the operation, and the sensitivity of selected measures is analyzed as a function of increased traffic volume. A selection of the analytical robustness measures listed in the previous section is compared and benchmarked against simulation-based measures through the generation of test timetables. The robustness of each individual timetable is described by compact indices, corresponding to either the analytical measure or to a descriptive index of the effects given by disturbances in a micro-simulated environment. These indices describe the increase of disturbance as a function of the primary delays, namely the

robustness. The metrics are, then, studied as a function of the number of scheduled trains to assess their sensitivity to the traffic volume. The measures identified are based on different scales. A share of the methods return a relative value in the range [0, 1] or a percentage, while the remainder measures return absolute values that require normalization for a comparison between different scenarios.

Being based on micro-simulation, the method is highly resources consuming, and a sampling procedure to reduce the number of iterations needed is proposed. Furthermore, the simulation process is simplified by the exclusion of dispatching strategies. This is a reasonable assumption on relatively short railway lines, where dispatching is usually realized at far ends rather than at intermediate stations. The simplistic First-In-First-Out rule at junctions can be considered a good approximation of real operation in these cases.

2.1.3.1 Robustness measures evaluated

The following analytical measures of robustness are investigated in this paper:

- Number of trains in the timetable (Traffic Volume) (Salido et al., 2008)
- Standard deviation of headways at a station (Carey, 1999)
- Mean absolute deviation of headways at a station (Carey, 1999)
- SSHR (Vromans et al., 2006)
- SAHR (Vromans et al., 2006)
- HETs (Haith et al., 2014)
- HETa (Haith et al., 2014)
- HetA (Landex and Jensen, 2013)
- Maximum Runtime Difference (Andersson et al., 2013a)

In addition, the following simulation-based measures are calculated as a benchmark:

- Capacity consumption (UIC, 2004)
- Total delay at selected stations (Solinen et al., 2017)
- Settling time (Salido et al., 2008)
- Number of delayed trains (Salido et al., 2008)
- Average delay per train (Salido et al., 2008)

Sensitivity to traffic volume increases

The analytical measures listed above provide compact indices that represent the reliability of timetables so the sensitivity to increased traffic volumes can be examined by normalization of the results.

Simulation-based measures of robustness, instead, describe the response of the railway system to given perturbations. Several simulations are required to describe the

change in the response as the effect of variations in the primary delays. Synthetic indices are here proposed to compare the results from simulation-based measures.

The relationship between a value of primary delay and the consequent total delay generated on the line was found polynomial of the second degree in analytical models (Landex, 2008) so the second derivative of a regressed polynomial from the simulation is proposed to address the amplification of the effect of disruptions. The total delay values measured for different primary delays can be regressed to a second-degree polynomial

$$d(p) = a \cdot p^2 + b \cdot p + c, \tag{11}$$

where p is the primary delay assigned to a train, and d is the measured total delay. The resulting total delay sensitivity index is so determined:

$$i_d = d''(p) = 2a. \tag{12}$$

Other simulation-based measures listed above have an irregular relationship to the primary delays. An analytical closed form the settling time, the number of trains delayed and the average delay per train is not available, so the evaluation of the amplification given by traffic volume increases can be operated by numerical minimization of the square distance between the measures in different timetable scenarios. An amplification factor is proposed here to compare irregular measures.

Define $S_t = \{s_{t_1}, s_{t_2}, \dots, s_{t_n}\}$ the array of settling time values measured on timetable t, after the generation of primary delays from 1 to n minutes. The original timetable is referred to as "a", and the measured array is $S_a = \{s_{a_1}, s_{a_2}, \dots, s_{a_n}\}$. Define the array $\overline{S_t} = m_t \cdot S_a = \{m_t \cdot s_{a_1}, m_t \cdot s_{a_1}, \dots, m_t \cdot s_{a_n}\}$. This is the curve to be associated to the timetable t, taking m_t as its multiplication factor. The amplification factor m_t is calculated minimizing the difference between the real measured curve and the multiplied one. This is each timetable's settling time indicator of sensitivity to primary delays.

$$\begin{aligned} m_t \coloneqq \min \left(\left(m_t \cdot p_{a_1} - p_{t_1}\right)^2 + \left(m_t \cdot p_{a_2} - p_{t_2}\right)^2 + \cdots + \left(m_t \cdot p_{a_n} - p_{t_n}\right)^2 \right) \\ = \sum_{j=1}^{n} \left(m_t \cdot p_{a_j} - p_{t_j}\right)^2 \end{aligned} \tag{13}$$

This approach is also valid for the number of trains delayed and the average delay per train as a function of the primary delay.

2.1.3.2 Micro-simulation and reduction of computational load

Railway microsimulation uses continuous computation of train motion equations and simulates the interaction between trains through discrete processing of signal boxes state. Given user defined infrastructure, rolling stock, and timetable databases, it is possible to calibrate the simulation through a performance parameter individually set for every train. This is a crucial parameter that influences the analyses output: it rules the percentage of train's maximum tractive effort used and the percentage of max allowed speed that the train will reach either in ordinary or delayed condition. Though it can reasonably be assumed that a delayed train driver tries to stick back to the timetable running at the maximum performance and speed available, it is hard to model the standard behavior. It is clear that higher performance parameter values for the standard operation reduce the running time margin, affecting the capability of one train to recover from delays along its path, increasing the follow-up delays.

The massive computation load of microsimulation is well known (Mattsson, 2007; Meester and Muns, 2007; Parbo et al., 2016). Therefore, a method is proposed here to reduce the number of scenarios to simulate and the resources needed, which we called the skimming method. The overall disturbance generated by a primary delay depends on the specific hindered train: according to its own scheduled running time supplement and to the margin time in the following train headway, the same disruption could affect different shares of trains and generate different amount of delays. For this reason, the disruption should be simulated against every train to measure its effect on the timetable, meaning considerable resources employment.

The skimming method proposed here consists of only applying a very detailed analysis of one parameter to the original timetable, measuring the effects of the same disruption given individually to each train. The analysis is not extended to all the trains, timetables and scenarios under test: it is rather the basis for the sampling in the search of the most representative train, with respect to the effect of disruptions. In order to contain the loss of information due to the reduction of simulation, an indicator of approximation goodness is introduced. The total delay is proposed to compare the impact of disruptions affecting different trains, as it synthetically represents the overall hindrance phenomenon through its magnitude.

The total delay on the line is measured as a function of primary delay separately for each train. The average total delay is then calculated among all the trains given a primary delay and choose the most representative one comparing its behavior with the

average. If we define the array of total delay values associated with each train primary delay from 1 to n minutes $D_c = \{d_{c_1}, d_{c_2}, \dots, d_{c_n}\}$ and the analogous average total delay array $\overline{D} = \{\bar{d}_1, \bar{d}_2, \dots, \bar{d}_n\} = \left\{\frac{\sum_{c=1}^{C} d_{c_1}}{C}, \frac{\sum_{c=2}^{C} d_{c_2}}{C}, \dots, \frac{\sum_{c=n}^{C} d_{c_n}}{C}\right\}$, the course c selected to be the most representative one is the one which total delay array is the closest to the average total delay $\overline{D}$:

$$c: \min\left(\left(\frac{\sum_{c=1}^{C} d_{c_1}}{C} - d_{c_1}\right)^2 + \left(\frac{\sum_{c=1}^{C} d_{c_2}}{C} - d_{c_2}\right)^2 + \cdots + \left(\frac{\sum_{c=1}^{C} d_{c_n}}{C} - d_{c_n}\right)^2\right) = \min \sum_{j=1}^{n}\left(\frac{\sum_{c=1}^{C} d_{c_j}}{C} - d_{c_j}\right)^2. \quad (14)$$

The load reduction can be estimated through the ratio between the number of simulations needed before and after the skimming method.

The array of the primary delay values generated is defined as $P_d = \{p_{d_1}, p_{d_2}, \dots, p_{d_n}\}$, where n is the number of primary delay values generated. In addition, the array of the traffic volumes of each test timetable, measured as the number of trains in the time period subject of study, is defined $V = \{v_1, v_2, \dots, v_{n_{tt}}\}$, where n_{tt} is the number of the test timetables.

The computational savings given by the skimming method identifies its efficiency comparing the numbers of simulation runs necessary with or without the sampling. Without the skimming method, a simulation run is theoretically necessary for every amount of primary delay, given to every train, on every scenario studied. The number of simulations needed is

$$n_{si} = n \cdot \sum_{j=1}^{n_{tt}} v_j \cdot n_{sc}, \quad (15)$$

with n_{sc} being the numbers of scenarios to be tested. In the skimming method, a detailed analysis of the original timetable is necessary on the original scenario, and a shrunk analysis of one delayed course for every timetable on each scenario. The number of simulations needed with the skimming method is

$$n_{si}^{*} = n \cdot v_1 + n \cdot n_{tt} \cdot n_{sc} = n(v_1 + n_{tt} \cdot n_{sc}). \quad (16)$$

The first addend refers to the detailed analysis to select the most representative train, while the second is the result of the shrunk analyses giving primary delay only to the selected train. The relative computation saving η is rated comparing the number of simulations needed in the two cases:

$$\eta = \frac{n_{si} - n_{si}^{*}}{n_{si}} = 1 - \frac{n_{si}^{*}}{n_{si}} = 1 - \left(\frac{v_1}{\sum_{j=1}^{m} v_j \cdot n_{sc}} + \frac{n_{tt}}{\sum_{j=1}^{m} v_j}\right) = 1 - \left(\frac{v_1}{\sum_{j=1}^{m} v_j \cdot n_{sc}} + \frac{1}{\overline{v_j}}\right), \quad (17)$$

where $\overline{v_j} = \frac{\sum_{j=1}^{m} v_j}{n_{tt}}$ is the average traffic volume per timetable.

Equation (17) shows an inverted hyperbolical saving as a function of the number of scenarios under test, meaning that the more scenarios, the better saving. The same relation is valid between the average traffic volume of the timetables and the saving, while it as a negative linear trend against the ratio between the original timetable and total traffic volume generated in all the timetables. In other words, the first term within the parenthesis quantifies the computational load of the first deep analysis to select a representative course compared to the total load of a complete analysis applied to every scenario. The second term quantifies the saving of the mere reduction in simulations needed.

2.1.3.3 Applicability

The method described, compares different scenarios simulating the railway system in its entirety. Different models of infrastructure, rolling stock, timetable, operation sets of rules, and signaling system can be tested and benchmarked. Furthermore, the effects of several modifications can be studied either individually or giving shape to combined changes to assess the joint benefits.

Different railway-like transport systems can be modeled in the micro-simulation tool, so the method can be applied, for instance, to metros, people movers, and other guided systems, making accuracy and flexibility the strengths of this method.

2.1.4 Application: the Oude Lijn in the Netherlands

The proposed method was used to evaluate the benefits of major works that are taking place on a Dutch densely occupied railway corridor. The current timetable runs 11 trains/h between The Hague and Rotterdam. The results are discussed in the following sections.

The railway is undergoing an infrastructure upgrade in Delft: a viaduct in the city center will be replaced by a tunnel. It is arranged to host four tracks, though the last two will be built in a second phase. The five set up scenarios represent respectively the current 2-tracked viaduct in Delft, the new railway tunnel in Delft with ceiling speed of through running trains increased from 100 km/h to 140 km/h, the planned four-tracked tunnel to Delft Zuid and a hypothetical extension of quadruple tracks to Rotterdam; besides, a signaling system modification is studied, being applied to the current viaduct infrastructure.

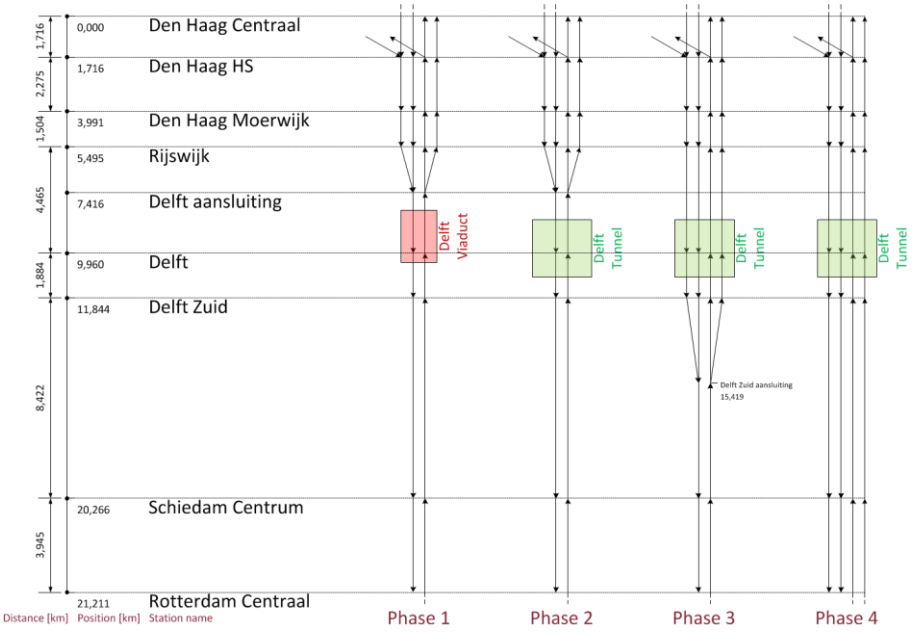

Figure 2.1-1: Track layout of each infrastructure scenario on the railway corridor

Scenario	Junction stationing	Tracks in Delft	Speed limit in Delft (km/h)	Status
Phase 1	North Delft	2 (Viaduct)	100	Current state
Phase 2	North Delft	2 (Tunnel)	140	Under construction (2015)
Phase 3	South Delft	4 (Tunnel)	140	Planned
Phase 4	None	4 (Tunnel)	140	Hypothetical
Signaling	North Delft	2 (Viaduct)	100	Hypothetical

Table 2.1-1: Infrastructure scenarios under comparison

The characteristics of the individual infrastructure scenarios are summarized in Table 2.1-1 and Figure 2.1-1. In this case study, we only considered the southbound traffic, pushing the northbound traffic and the interaction between opposite flows to further studies.

The case study is based on the microsimulation tool OpenTrack (Nash and Huerlimann, 2004).

2.1.4.1 Traffic volume

The sensitivity of the robustness to increases in traffic volume is under examination, so additional timetables were developed starting from the original one, with

limited modifications at the train characteristics. Stopping patterns and order of trains are maintained for all the trains and new trains scheduled should only be copies of existing trains in the timetable.

Starting from an original timetable, which is the reference in the following steps, new timetables were developed increasing the traffic volume. The test-timetables were built by stepwise frequency increases scheduling additional services to the reference timetable, keeping the existing trains order.

The original timetable may include different categories of trains. The share of train categories within a timetable is one of its peculiar characteristics. For this reason, the ratio between the size of the categories should be kept equal, or at least on the same scale, in all the timetables. Defined the array containing the number of trains of each category in the timetable t, $C_t = \{c_{t_1}, c_{t_2}, \dots, c_{t_z}\}$, and the total traffic volume of each timetable

$$v_t = \sum_{l=1}^{z} c_{t_l}, \tag{18}$$

the share $\frac{c_{t_q}}{\sum_{l=1}^{z} c_{t_l}} = \frac{c_{t_q}}{v_t}$ is maintained through the timetables t for each category q in the total traffic volume.

The traffic volume can be increased up to consuming the whole capacity, in which case the buffer times between services are nulled and the running time supplements are reduced to shrink the heterogeneity. In fact, the slower trains are speeded up in the schedules, and their running time supplement is reduced to the minimum to reduce the line exploitation. According to Huisman and Boucherie (2001), the maximum capacity of a railway line corresponds to all equal train paths.

A total of five timetables were built to test the infrastructures: the current timetable with 11 trains/h was called "A" and the traffic volume was increased stepwise up to 18 trains/h in timetable "E". The performance parameter was updated in timetables D and E to fit shorter running times for local trains and reach better homogeneity among train paths; in every case the minimum time supplement was satisfied. A set of integer values of primary delays is selected in the range from 1 to 10 minutes: $P_d = \{1\ min, 2\ min, \dots, 10\ min\}$, with $n = 10$: these can be considered typical daily disruptions, due to, among others, boarding at stations, minor failures at the rolling stock or at the infrastructure.

The skimming method was applied to reduce the number of simulations needed. The original timetable was timetable A, and the reference infrastructure scenario was Phase 1, with the following results:

$$n_{si} = 3550, n_{si}{}^{*} = 250, \eta = 89{,}86\%.$$

The measured total delay resulting from primary delays given to every train is shown in the graph below: the average curves and the curve of the most representative course are highlighted.

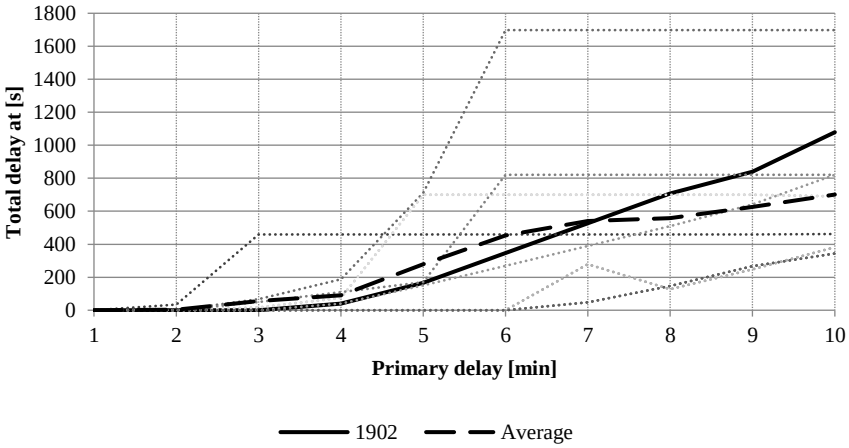

Figure 2.1-2: Total delay resulting from each trains' primary delay. The average and the selected train's total delay curves are highlighted

The resulting robustness indices are summarized in the table below for every infrastructure scenario. HETs and HETa are calculated with reference to a nominal headway of 1 minute. All the measures specific to one station were calculated for arrivals at Rotterdam Central station.

The operation of the 2-tracked tunnel will permit an increase of the train frequency from 11 trains/h up to 12 trains/h per direction, while the extension of quadruple tracks from Rijswijk through Delft will enable the operation of up to 16 trains/h.

The maximum UIC 406 leaflet track capacity consumption will be reduced from currently 80% to 70% and finally 60% for the different infrastructure and basic timetable scenarios. The total delay of southbound trains as a function of the primary delay is well described by quadratic parabolic functions. The sensitivity of the infrastructure and timetable to knock-on delays does not change significantly after the operation of 12 trains/h per direction on initially 2 tracks in the new tunnel in Delft.

Infrastructure scenario	Timetable j	A	B	C	D	E
	Traffic volume v_j (trains/h)	**11**	**12**	**14**	**16**	**18**
All	Standard deviation headways	0,49	0,50	0,34	0,40	0,40
	M.a.d. headways	0,78	0,80	0,73	0,75	0,73
	SSHR*	1,00	1,15	1,57	1,77	2,23
	SAHR*	1,00	1,15	2,07	2,16	2,73
	HETs	0,51	0,53	0,62	0,61	0,69
	HETa	0,23	0,24	0,38	0,34	0,39
	Het_A	0,49	0,46	0,59	0,67	0,72
	MRD*	1,00	1,00	1,00	0,93	0,87
Phase 1	Capacity consumption	80,8%	86,1%	98,9%	98,9%	100,0%
	Average running time margin	21,6%	22,2%	20,4%	19,6%	18,0%
	Total delay*	1,00	1,06	1,70	2,30	3,38
	Settling time*	1,00	0,99	1,16	1,21	1,71
	Number of delayed trains*	1,00	1,00	1,39	1,64	2,54
	Average delay per train*	1,00	0,98	0,87	0,90	0,96
Phase 2	Capacity consumption	70,3%	71,9%	89,4%	88,1%	96,7%
	Average running time margin	26,0%	27,0%	25,1%	23,9%	22,1%
	Total delay*	0,95	1,05	1,61	2,27	2,58
	Settling time*	1,00	0,98	1,14	1,19	1,61
	Number of delayed trains*	1,00	1,00	1,35	1,56	2,35
	Average delay per train*	0,84	0,80	0,71	0,72	0,78
Phase 3	Capacity consumption	58,6%	60,3%	76,1%	76,1%	86,1%
	Average running time margin	26,8%	27,8%	25,7%	24,6%	22,9%
	Total delay*	0,96	1,05	1,59	1,84	1,88
	Settling time*	1,00	0,98	1,15	1,19	1,33
	Number of delayed trains*	1,00	1,00	1,35	1,52	1,68
	Average delay per train*	0,84	0,81	0,72	0,69	0,89
Phase 4	Capacity consumption	60,8%	63,1%	70,6%	73,9%	78,3%
	Average running time margin	27,9%	29,0%	26,8%	25,6%	23,9%
	Total delay*	1,11	1,16	1,52	1,82	2,26
	Settling time*	1,00	0,99	1,12	1,16	1,37
	Number of delayed trains*	1,00	1,00	1,34	1,52	1,41
	Average delay per train*	0,77	0,75	0,68	0,63	0,79
Signaling	Capacity consumption	60,3%	65,0%	83,3%	82,5%	88,9%
	Average running time margin	23,6%	24,3%	22,4%	21,5%	19,9%
	Total delay*	1,04	1,07	1,61	1,85	2,20
	Settling time*	1,00	0,99	1,17	1,20	1,32
	Number of delayed trains*	1,00	1,00	1,53	1,74	2,00
	Average delay per train*	1,02	0,98	0,84	0,83	0,76

Table 2.1-2: Calculation of performance indices for the five infrastructure scenarios. Asterisks identify the normalized measures

The measures of dispersion of the headways remain approximately unchanged across different timetables, despite the increased traffic volume. In particular, the standard deviation of headways and the mean absolute deviation are not affected by changes in the number of trains.

In all the scenarios, the total delay results being the most sensible measure to increases of traffic volume.

2.1.4.2 Discussion

The skimming method allowed a reduction of the workload of almost 90%, opening the possibility of a dramatic reduction of computation time and resources needed. The assumed absence of dispatching only affected few simulations with order alterations in the 2-tracked sections of the line. Considerable impact on the performance came, instead, from dispatching at the final station, with particular regard to the Phase 4 scenario. The track layout and allocation within the station generated several interdependencies and itinerary conflicts between arriving trains, resulting in queuing on the open line and in worse overall robustness. The capacity at stations is a known complex problem (Landex and Jensen, 2013), and the consequences on lines and on networks should not be disregarded. In this particular case, in the Phase 4 scenario, Rotterdam central station would be the bottleneck, being a critical robustness sink.

The results highlight the different sensitivity of the parameters to primary delays and to traffic volume. In particular, the measures of headway dispersion are not sensitive to increases in traffic volume and only account for heterogeneity of headways. This is predictable, as these measures are normalized to the total cycle time. Potentially, this means that a timetable with lower capacity consumption but uneven allocation of train paths over time might result in worse score than a homogeneous timetable that saturates the line. The capacity consumption seems to be quite an incomplete indicator for robustness: it is affected by the performance parameter and the running time supplements. In all the scenarios the capacity consumption did not increase from a traffic volume of 14 trains/h to 16 trains/h and in some cases, it decreased; the same happened for the last traffic volume boost, showing restricted increases. This is due to the change of the performance parameter in the schedule: trains can easily be compressed together with tighter schedules, and no evidence in the capacity consumption, as shown by timetables C, D, and E; on the other hand, the time supplement availability to recover from delays shrinks to the minimum, and disruptions' effects grow. In the specific case study, the best reduction in capacity consumption was observable in Phase 3, which was the only one to tolerate up to

16 trains/h keeping safe in 75 % limitation of peak hour capacity utilization (UIC, 2004). This phenomenon is much related to the relation between MRD and the traffic volume. In fact, MRD does not change until the slower trains are speeded up to homogenize the train paths. As a result, MRD cannot represent robustness without the support of other measures, as reductions in scheduled running times are often related to higher capacity consumption, meaning a higher risk of delay propagation, and effects opposed to the measure direction.

The total delay was the parameter most impacted by traffic volumes increase: for the current infrastructure – Phase 1 – its sensitivity to primary delays more than triplicated comparing a timetable witch 18 trains/h with the current 11 trains/h. It is also noticeable that sensible differences between the simulated scenarios could only be appreciated with 16 trains/h or more.

The settling time showed more stability against the traffic volume increase: its sensitivity index kept below 1,25 up to 16 trains/h. In addition, it seemed more independent on the upgraded infrastructure, as all the scenario gave very similar sensitivity values up to 16 trains/h.

Surprisingly, the share of trains involved in the disruption appeared remarkably stable with small traffic growth. From 14 trains/h, though, the sensibility spread more than the other indices, up to rather high values. This could be explained by the nature of the case study: the main differences between infrastructure scenarios were after the point of disruption, and delays were measured at the end of the line; the timetables A and B were almost identical, and it is understandable that infrastructure improvement like the partial extension of a 4-tracked stretch would not enhance the ability of single trains to recover from their own delay. At the same time, the average delay per train seemed totally blind to the traffic volume. Every scenario showed that the grade of the regressed line would slightly drop in higher traffic volumes. Moreover, the 0-delay point of the regressed line was linked to lower primary delay values when the traffic increased. This can be interpreted as a rise of interaction between trains with the traffic increase so that more trains are hindered by the previous train but in a smaller amount.

Moreover, the examination of different infrastructure scenarios highlights that simulation-based measures better show the interaction between schedules and the other components of a transport system, which are the rolling stock and the infrastructure. Indeed, the benefits of upgraded infrastructure are hidden in the synthetic analytical measures presented. The measures of heterogeneity in headways proposed by Haith et al. (2014) take into account this aspect, although this is only done considering a reference

minimum headway between trains, which does not account for possible different itineraries on the open line or within the stations.

According to the results collected, it could be stated that the robustness of operation will not be improved by the new tunnel in Delft until it is provided with four tracks. Good results were also reached through the bare signaling system upgrade on the current infrastructure: sensitivity reduction of parameters against primary delay could be obtained through this measure, measurable similar to the advantage given by the 4-tracked section extension. Furthermore, comparison of all the indicators' behavior as functions of frequency in all the infrastructures showed that the 4 tracks section extension would reduce the Total Delay by reducing the number of trains hindered, while a closer interaction between the trains would be gained by the signaling works, hindering more trains by a lower amount. By implementing ETCS Level 1 with braking curve supervision instead of changing track layout, the robustness indicators would improve as in the scenario with the extension of the complete line with 4 tracks to Delft Zuid without ETCS.

Gathering the information from the indices we could state that rising the traffic volume, the settling time seems to be marginally affected, while the total delay raises. It means that the disruption should take effect within the same lapse of time. At the same time, the average delay per train reveals as independent from the increase of traffic volume, which means that disruptions spread among trains in packed timetables, rather than increasing the amount individual delay; the two pieces of information match indeed, meaning that the total delay increases because more trains are contained in the same settling time, each of them is hindered by the same amount.

2.1.5 Conclusions and further studies

This paper presented an effective and economical method to benchmark infrastructural and operational scenarios. The method proposed highlighted the necessity of a further step in new infrastructure building in Delft: real benefits to robustness will be achieved only by the extension of the 4-tracked section. Similar results could be obtained by new a signaling system implementation, although it would be not feasible for just one line on the network.

The method allowed the comparison of different infrastructure scenarios and showed the efficacy of some actions to improve operation stability, rather than others. The flexibility given by a micro-simulation based approach suits to benchmark and compare different infrastructures, rolling stocks operational rules and railway-like transport system. In addition, an effective procedure is proposed to reduce the heavy computational load

typical of microsimulation to make the analyses lean; the skimming method could be improved and adapted to other contexts to reduce the computing time, which would open the gates to the use of micro-simulation in real-time problem solving such us re-scheduling.

Different robustness measures were compared in this paper, with a particular focus on their sensitivity to traffic increases and their ability to represent the loss of reliability associated with more intense exploitation of the infrastructure. Measures based on simulation, representing the overall disturbance on the operations revealed to be more accurate than synthetic measures based mainly on the heterogeneity in the timetable.

This research's implications include the availability of new negotiation tools between Infrastructure Managers and Railway Undertaking; the benchmarking is needed to measure improvements from different solutions. The paper shows the lack of information of the capacity consumption indicator about robustness, even though a correlation between the line exploitation and timetable's robustness evidently exist. The relation should be examined in depth together with the link between capacity consumption, headways between trains, running time supplements and robustness.

ACKNOWLEDGMENT: The data collection for this research was funded by the Sapienza University of Rome and the Erasmus Program of the European Union; the research was also supported by the Delft University of Technology.

2.1.6 References

Andersson, E., Peterson, A., Törnquist, J., 2013a. Quantifying railway timetable robustness in critical points. J. Rail Transp. Plan. Manag. 3, 95–110. doi:10.1016/j.jrtpm.2013.12.002

Andersson, E., Peterson, A., Törnquist Krasemann, J., 2013b. Introducing a New Quantitative Measure of Railway Timetable Robustness Based on Critical Points, in: Proceedings of 5th International Seminar on Railway Operations Modelling and Analysis (IAROR): RailCopenhagen2013. Copenhagen, pp. 1–19.

Andersson, E., Peterson, A., Törnquist Krasemann, J., 2011. Robustness in Swedish Railway Traffic Timetables, in: Ricci, S., Hansen, I.A., Longo, G.L., Pacciarelli, D., Rodriguez, J., Wendler, E. (Eds.), Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis. Rome, pp. 1–18.

Barron, A., Melo, P., Cohen, J., Anderson, R., 2013. Passenger-Focused Management Approach to Measurement of Train Delay Impacts, in: Transportation Research Board, 92nd Annual Meeting. Transportation Research Board of the National Academies, pp. 46–53. doi:10.3141/2351-06

Carey, M., 1999. Ex ante heuristic measures of schedule reliability. Transp. Res. Part B Methodol. 33, 473–494. doi:10.1016/S0191-2615(99)00002-8

Carey, M., Kwieciński, A., 1994. Stochastic approximation to the effects of headways on

knock-on delays of trains. Transp. Res. Part B 28, 251–267. doi:10.1016/0191-2615(94)90001-9

Corman, F., D'Ariano, A., Hansen, I.A., 2014. Evaluating disturbance robustness of railway schedules. J. Intell. Transp. Syst. Technol. Planning, Oper. 18, 106–120. doi:10.1080/15472450.2013.801714

Gibson, S., Cooper, G., Ball, B., 2002. Developments in transport policy: The evolution of capacity charges on the UK rail network. J. Transp. Econ. Policy 36, 341–354.

Ginkel, A., Schobel, A., 2007. To Wait or Not to Wait? The Bicriteria Delay Management Problem in Public Transportation. Transp. Sci. 41, 527–538. doi:10.1287/trsc.1070.0212

Goverde, R.M.P., Hansen, I.A., 2013. Performance indicators for railway timetables, in: 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings. IEEE, pp. 301–306. doi:10.1109/ICIRT.2013.6696312

Haith, J., Johnson, D., Nash, C., 2014. The case for space: the measurement of capacity utilisation, its relationship with reactionary delay and the calculation of the capacity charge for the British rail network. Transp. Plan. Technol. 37, 20–37. doi:10.1080/03081060.2013.844906

Harker, P.T., Hong, S., 1994. Pricing of track time in railroad operations: An internal market approach. Transp. Res. Part B 28, 197–212. doi:10.1016/0191-2615(94)90007-8

Hofman, M.A., Madsen, L., Groth, J.J., Clausen, J., Larsen, J., 2006. Robustness and Recovery in Train Scheduling - a simulation study from DSB S-tog a / s. 6th Work. Algorithmic Methods Model. Optim. Railw. 97–118. doi:10.4230/OASIcs.ATMOS.2006.687

Huisman, T., Boucherie, R.J., 2001. Running times on railway sections with heterogeneous train traffic. Transp. Res. Part B Methodol. 35, 271–292. doi:10.1016/S0191-2615(99)00051-X

Jensen, L.W., 2015. Robustness indicators and capacity models for railway networks. Technical University of Denmark.

Landex, A. (2008). Methods to estimate railway capacity and passenger delays. Technical University of Denmark (DTU). Retrieved from http://findit.dtu.dk/en/catalog/2185768953

Landex, A., Jensen, L.W., 2013. Measures for track complexity and robustness of operation at stations. J. Rail Transp. Plan. Manag. 3, 22–35. doi:10.1016/j.jrtpm.2013.10.003

Mattsson, L.-G., 2007. Railway Capacity and Train Delay Relationships. Crit. Infrastruct. Adv. Spat. Sci. doi:10.1007/978-3-540-68056-7_7

Meester, L.E., Muns, S., 2007. Stochastic delay propagation in railway networks and phase-type distributions. Transp. Res. Part B Methodol. 41, 218–230. doi:10.1016/j.trb.2006.02.007

Nash, A., Huerlimann, D., 2004. Railroad simulation using OpenTrack. Comput. Railw. IX 45–54.

Parbo, J., Nielsen, O.A., Prato, C.G., 2016. Passenger Perspectives in Railway

Timetabling: A Literature Review. Transp. Rev. 36, 500–526. doi:10.1080/01441647.2015.1113574

Parbo, J., Nielsen, O.A., Prato, C.G., 2014. User perspectives in public transport timetable optimisation. Transp. Res. Part C Emerg. Technol. 48, 269–284. doi:10.1016/j.trc.2014.09.005

Peterson, A., 2012. Towards a robust traffic timetable for the Swedish Southern Mainline, in: Computers in Railways XIII. WIT Transactions on The Built Environment, New Forest, UK, pp. 473–484. doi:10.2495/CR120401

Salido, M.A., Barber, F., Ingolotti, L., 2008. Robustness in railway transportation scheduling, in: Proceedings of the World Congress on Intelligent Control and Automation (WCICA). IEEE, Chongqing, China, pp. 2833–2837. doi:10.1109/WCICA.2008.4594481

Schittenhelm, B.H., 2011. Planning With Timetable Supplements in Railway Timetables, in: Annual Transport Conference at Aalborg University. trafikdage, Aalborg, DK.

Solinen, E., Nicholson, G., Peterson, A., 2017. A microscopic evaluation of railway timetable robustness and critical points. J. Rail Transp. Plan. Manag. 7, 207–223. doi:10.1016/j.jrtpm.2017.08.005

UIC, 2004. Leaflet 406 - Capacity.

Vromans, M., Dekker, R., Kroon, L.G., 2006. Reliability and heterogeneity of railway services. Eur. J. Oper. Res. 172, 647–665. doi:10.1016/j.ejor.2004.10.010

Wiklund, M., 2002. The vulnerability of the railway transport system - A structure for formulation of models and development of methods, VTI meddelande. Linköping. doi:0347-6049

3 AN ANALYTICAL DELAY PROPAGATION MODEL

3.1 Paper II: A Closed Form Railway Line Delay Propagation Model

Cerreto, Fabrizio, Steven Harrod, and Otto Anker Nielsen. "A Closed Form Railway Line Delay Propagation Model." Submitted to *Transportation Research Part C: Emerging Technologies*, October 24, 2017. Re-submitted after the second round of review, February 3, 2018

Presented at the *6th European Transport Research Conference* (Transport Research Arena), Warsaw, Poland, April 18-21, 2016, at the annual meeting of the *Institute for Operations Research and the Management Sciences* (INFORMS 2016), Nashville, USA, November 12-16, 2016, and at the *Transport Conference* (Trafikdage 2017), Aalborg, Denmark, August 29-30, 2017.

Abstract

Railway service quality can be measured by the aggregate delay over a time horizon due to an event that delays a given train. Timetables for railway services may dampen delay propagations to subsequent trains by adding either supplement time or buffer time to the minimum driving time. The evaluation of these variables is often performed by time-consuming analysis with simulation software. This paper proposes instead an analytical closed-form formulation of aggregate delay. This can be used to obtain theoretical insights into railway delays and as a component of larger railway scheduling models, where the iterative use of simulation models would require far too much calculation time. Analysis of the function recommends a slack control policy, as the delay-damping effect of supplement and buffer decreases with their magnitude. Further, the effect of different threshold values in delay measurement is demonstrated, giving information valuable to the design of service contracts. Numerical analysis of a railway line in Copenhagen shows that the polynomial function provides guidance and insight even when theoretical assumptions are violated.

KEYWORDS: *Rail transportation, Train delays, Timetable robustness, Timetable design, Delay propagation*

3.1.1 Introduction

Operational stability and robustness are important for railway transport. Not only are the passengers sensitive to these measures of quality (Parbo et al., 2016), but railways are usually integrated networks, where failures at one location often affect other locations and services. Railway network planners are thus faced with many decisions about what quality of service to provide and what resources to allocate to deliver this service. Much of the literature demonstrates that there are often multiple feasible alternatives to allocate timetable allowance, and each alternative has a unique performance profile with regard to punctuality and robustness (Caimi et al., 2009). The analysis of these alternatives frequently requires laborious modeling with simulation software, which is time-consuming in both model programming and analysis run-time (Carey, 1999; Carey and Kwieciński, 1994; Parbo et al., 2014). Faster and simpler methods for performance appraisal use stochastic simulation models or analytical approaches, the former being more suitable when the timetable is unknown, and the latter being able to include a deeper level of detail (Meester and Muns, 2007). Analytic models are known to be much faster than simulation models, though the former require simplifying assumptions that might lead to inaccurate results (Mattsson, 2007).

This paper contributes to the literature with an analytic closed form formulation of aggregate railway line delay propagation in response to a primary delay. This function may supplement or replace the application of simulation models for exploration of alternatives when appraising different timetable alternatives. The formulation is closed form under a set of timetable-structure assumptions. It is later shown in this paper, using microsimulation, that the formulation is robust to deviation from these assumptions. The mathematical model facilitates a quick evaluation of the expected cumulative delay and makes it possible to evaluate structural design factors, such as running time supplement, headway buffers, and to design service contract performance measures. Lastly, cumulative delay calculated in a closed form can efficiently be implemented in optimization models for timetabling.

The formulation is derived from a finite series of deviations from the service plan (secondary delays) caused by a singular initial disruption (primary delay). The primary delay is propagated to following trains and recovered on the individual trains' downstream paths. The scientific novelty of this model is the explicit and simultaneous inclusion of headway buffers and running time supplements to reduce the individual train delay in the propagation process, whereas previously proposed methods considered only one type of

timetable slack at a time, or used queuing theory to evaluate the interferences between vehicles (Hasegawa et al., 1981; Huisman, 2002; Landex, 2008; Mattsson, 2007; Pyrgiotis, 2012; Salido et al., 2012). Furthermore, it is possible to model multiple primary delays and to evaluate the response on railway lines and networks, with the possibility to disregard small delays under a defined threshold.

3.1.2 Literature Review

3.1.2.1 Aggregate delay measures

Cumulative, aggregate, or total delay, are common performance measures used in several fields, from operations monitoring, to timetable planning and optimization. Academic research and industrial applications show the relevance of such metrics in timetable planning and management.

Following a comparison of methods and data used to assess performance quality by 22 metro operators worldwide, Barron et al. (2013) describe measures of the total effect of disruptions as the best representation of service quality, as they provide a better understanding of how incidents affect operation and customers. In particular, total vehicle hours of delay reflect the operator's interest in the vulnerability to network disruptions. Different forms of aggregate delay are currently used in operation analysis in the transport industry for service quality assessment, and to enforce contracts between railway undertakers and infrastructure managers. Transport for London uses Lost Customer Hours as a performance measure in metro operation. The measure consists of the total delay given to passengers, counted as estimated travel time extension due to incidents (TfL Investment Programme Management Office, 2008). In Europe, the Performance Regime produces cash flows between railway undertakers and infrastructure managers as an incentive to improve service quality. In Italy, every minute of train delay is valued at 2€ (Rete Ferroviaria Italiana, 2015), and the aggregate line delay can be correlated to the total cash flow generated by an individual primary delay. Diverse ways to extract this performance measure from past data, and to identify the main influencing factors have been proposed. Using regression analysis on data recorded in the British railway network, Gibson et al. (2002) identify an exponential functional relation between the line capacity utilization and the expected reactionary delays on a railway line. They furthermore identify several factors influencing the expected reactionary delays: geography, time of operation, and speed heterogeneity. Goverde and Meng (2011) developed a data analysis tool to build conflict trees based historical data on track occupation from the Dutch railways. The conflict trees depict the realized delay propagation across consecutive trains on a railway infrastructure,

so that it is possible to identify primary delays and the overall disturbance generated in form of secondary delays. Goverde and Meng assess the severity of individual disturbances by measuring the total delay that they generate.

The cumulative delay is also used as a metric in the planning phase to evaluate timetables before the real operation. It is often compared to given initial delays to evaluate how stable the timetable is against disturbances. Landex (2008) uses total delay as a measure of timetable reliability in his analytical model, and to define a relation between capacity consumption and total delay generated by a given single primary delay. Salido et al. (2012) compare timetables using the cumulative delay resulting from simulation and define a timetable more robust than another, if, for a given disturbance, the cumulative delay generated is smaller. Cerreto (2015) introduces a method to reduce computation time in simulation models, shrinking the number of simulation runs required with a heuristic process called the *skimming method*. The method is based on the measure of aggregate line delay in perturbed simulation scenarios. A composite profile of aggregate line delay is estimated from an initial simulation analysis.

The cumulative delay is also found strongly correlated with other performance measures, which makes cumulative delay a valuable objective candidate for timetabling optimization problems (Toletti, 2016; Törnquist, 2007). In delay management problems, Ginkel and Schobel (2007) optimize the operational departure time for a connecting service at a transfer station, given that the primary service is delayed. The aggregate delay is incorporated in the optimization function of a bicriteria model that minimizes the number of passenger missed connections and the total delay recorded by vehicles. Harker and Hong (1994) use cumulative line delay to evaluate dispatching choices on a railway network in a Nash noncooperative game. The model is set up to seek the network optimal dispatching strategy, given that the single divisions of the network act to minimize the aggregate deviation in their own area of control. The model is eventually used in the pricing of train slots, according to the value of time attributed to individual trains.

3.1.2.2 Delay propagation models

The interferences between trains under perturbed operation are expressed by delay propagation models. These models seek to mimic the development of secondary delays when primary delays are known. In this section, a survey of existing delay propagation models is provided, divided by models for railway lines and models for railway networks.

Models for railway lines

The methods listed in this section model delay propagation on railway lines, mainly unidirectional. Although these methods might appear alike, they differ by the input and output variables, and by the modeled interaction between trains.

Hasegawa et al. (1981) borrows concepts from road transport and applies a hydrodynamic analogy to model railway traffic. The study models the delay propagation on a unidirectional railway line as a shockwave in a compressible fluid. Timetable slack is modeled implicitly through speed and flow, where recovery is provided by trains running at higher speed and flow than scheduled. Discontinuities of traffic flow and density propagate at a speed that is independent of the entity of primary delays. The total delay is calculated as the integral of individual train delays in the domain of space and time, resulting as a cubic function of the primary delays. The model relies on measures hard to calculate in the planning process, such as spatial density, recovery flow and speed, and requires simulation for parameter calibration.

Carey and Kwiecinski (1994) propose a stochastic approximation of the secondary delays of trains. Realized trip times on a line segment are derived from distributions of primary delays and headway buffers. The study finds that when trains have exponential delays between stations, the expected trip time between stations is directly dependent on the headway between trains plus a constant.

Carey (1999) provides a theoretical description of the process delays transfer between consecutive trains at one station, and calculates the expected individual train delays, given the delay distributions of single trains and the set of headways at the station. The gain in reliability given by marginal timetable slack fades out with its magnitude under the assumption of downward sloping delay distributions. Carey's metrics refer to a single station, and individual train recovery along the path is not considered.

Huisman and Boucherie (2001) model the delay propagation in absence of a timetable, assuming that all trains run at their maximum possible speed. The first train is assumed to run within its minimum running time, whereas the following trains increase their running time in terms of the delay of the previous train, reduced by the headway buffer. Using queuing theory, expected interference between consecutive trains are estimated, but the influence of primary delay is only considered implicitly as the distribution of free-flow running times.

Mattsson (2007) offers a literature survey on reliability measures of railway services and the relationship between capacity consumption and unreliability. Mattsson

models the expected transit time over a line section as the summation of a minimum running time and a stochastic extension. Delay recovery is modeled through a share of expected delay included in the schedule. The objective of railway planners is the minimum expected passenger loss, which combines the scheduled running times and the expected unscheduled delays. Besides, Mattsson applies again an analogy to standard road traffic-flow to calculate the capacity utilization of a double-tracked railway line, as the percentage of time used by a train sequence. The method is similar to previous approaches (Gibson et al., 2002; UIC, 2004), and the capacity used by a timetable has been used later as a measure of its reliability (Haith et al., 2014).

Lastly, and most closely related to this paper, Landex (2008) proposes a delay propagation model computing the transfer of delay between trains through the scheduled buffer times. This model is used to study the relationship between capacity consumption and the development of the disruptions but does not consider the recovery of train delays according to the running time supplement. Landex hypothesizes homogeneous traffic on a single railway line and proposes timetable slack aggregation to model heterogeneous train paths, using average buffer time between pairs of trains. The method is later integrated with running time supplement by Jensen et al. (2017), who estimate the capacity consumption of a timetable under stochastic sets of primary delays in a mesoscopic simulation framework.

Models for railway networks

The interaction between trains on the railway network is more complex than railway lines. Different lines merge and diverge at stations, service constraints are introduced to satisfy the connections between trains and meet and passes are often scheduled at stations. Timed graph events are often utilized to represent these complex connections. The methods listed below differ mainly in the design of dependencies between scheduled events.

Zhu and Schnieder (2000) propose a model to assess railway timetable performance using Colored Petri Nets. Given primary delay distribution, the authors evaluate the timetable performance measuring the Delay Degree, which is the aggregate delay recorded at specific stations. Timetable planners can improve the performance changing the sequence of headways at individual stations. The complexity of Colored Petri Net models increases exponentially with the number of stations, and the authors point out the imperative need to reduce the model complexity.

Meester and Muns (2007) model the realized process times in a stochastic timed event graph as a combination of a minimum process time and a random extension. The initial delay of a process is recovered by a process time supplement, and the final residual delay is transferred to the downstream events. Delays are propagated recursively in a continuous Markov chain, given the distribution of process time extension for every connection between arrival and departure events. The paper states that a phase-type distribution, a distribution of the absorption time of a continuous time Markov chain, can be contained in closed form using three operations on individual delays: sum, nonnegative excess beyond a bound, and maximum. The method requires knowledge on the distribution of primary delays for individual processes and depends on the assertion of independence of the primary delays. Resembling simulation models, it is not possible to extract a functional relationship between primary and secondary delays.

Goverde (2010, 2007) presents an efficient delay propagation algorithm where timetables are modeled as timed event graphs (using max-plus algebra) and initial delays are known. The algorithm is very fast and, in a few seconds, can calculate the delay propagation over a large network consisting of many interdependent services, such as the Dutch national railway timetable. Goverde uses this method to compute performance indicators, including delay propagation statistics such us total secondary delay, and settling time. However, the model offers no functional relationship, and results must be calculated for each scenario separately.

Graph methods for delay propagation occur also in other means of transportation. Pyrgiotis (2012) describes a mixed algorithmic and analytical model to propagate delays in airport networks, following aircraft rosters, based on queuing theory. The analytical section of this model uses queueing theory to assign delays to flights due to congestion at airports, based on the airport capacity and the time-dependent traffic demand. The delay assigned to a flight is then propagated to the downstream airports in the aircraft's roster. The aircraft delay is reduced in every trip segment by means of the scheduled slack, and it is increased by congestion, from queuing theory. Table 3.1-1 summarizes the mentioned references and offers a comparison of their main features.

The applications listed in section 3.1.2.1 reveal the significance of aggregate delay as a performance measure, both in research and in industry. In the successive sections, the literature on delay propagation models and aggregate delay estimation is reviewed.

PAPER	APPROACH	INFRASTR.	TRAFFIC	TIMETABLE	CONTROL PARAMETERS	INPUT	OUTPUT	FUNCTI. REL.
Hasegawa, 1981	Analytical	L	U, H	K	Design values and maximum values for speed, flow, density	Primary delay	Aggregate line delay	Y
Carey and Kwiecinski, 1994	Stochastic	L	U	K	Scheduled Headways	Distributions of Free and Minimum running times	Individual trip times	Y
Carey, 1999	Stochastic	S	A	U	Headway buffers	Distributions of primary delays	Individual expected delays	N
Huisman and Boucherie, 2001	Stochastic	L	U	U	Scheduled headways	Distributions od free running time, Distributions of actual headway	Distributions of actual running time	N
Mattsson, 2007	Analytical	L	A	U	Running time supplement	Distribution of primary delays	Expected running time	N
Landex, 2008	Analytical	L	A	K	Headway buffers	Primary delay	Aggregate line delay	Y
Zhu and Schnieder, 2000	Simulation - Colored Petri net	N	A	K	Scheduled Headways	Distributions of primary delays	Aggregate station delay	N
Meester and Muns, 2007	Analytical	N	A	K	Process time supplement	Distributions of process times	Expected delays	N
Goverde, 2007,2010	Analytical	N	A	K	Running time supplements, Headway buffers, Connection buffers	Primary delay, Minimum headways, Minimum dwell times, Connections	Total secondary delay, number of delayed trains, average secondary delay, settling time	N
Pyrgiotis, 2012	Stochastic + Analytical	N	A	K	Running time supplement	Airport capacity	Individual delays	N
Cerreto et al., 2018 (this paper)	*Analytical*	*L, N**	*U, H*	*K*	*Running time supplement, Headway buffer, Delay threshold*	*Primary delay*	*Aggregate line delay, individual train delays, settling time*	*Y*

*Table 3.1-1 Literature summary. Infrastructure: N – Network, L – Line, S – single station, *With recursive application. Traffic: A – Any direction, U – Unidirectional, H – Homogeneous. Timetable: K – known, U – unknown.*

Most of the delay propagation models for railway lines do not consider the recovery through the running time supplements. In a few cases, running time supplements are included, but delay recovery through the headway buffer between trains is disregarded. Even though many of these models provide insight into the interferences between delayed trains, they are very theoretical, and their application appears limited to railway lines and to single primary delays. Moreover, the only closed form function that returns the cumulative delay generated by a given primary delay, which considers recovery both between trains and along the train path is given by Hasegawa (1981), but this model does not consider conditions of partial delay recovery or no recovery, and uses control parameters that are difficult to measure and calibrate.

More complex models based on event graphs are able to include both the types of recovery at the same time and to represent the multiple dependencies between timed events on railway networks. These models, though, do not provide a functional relationship between primary delays and aggregate delay, so the insight on the relation is limited to a test-based analysis.

Furthermore, very seldom in the presented literature, a delay threshold is considered to disregard small delays, although it is a common tuning parameter for performance assessment (European Performance Regime project, 2013; Hofman et al., 2006; Jensen, 2015; Landex, 2008; Parbo et al., 2016; Schittenhelm, 2013, 2011; UIC, 2009; Vromans, 2005).

For this reason, the model presented in this paper is designed as a closed form function of primary delays and aggregate delays, where recovery is possible between both trains and stations. The model represents diverse recovery conditions, and supports multiple primary delays, fits railway networks, and includes a delay threshold under which delays can be disregarded.

3.1.3 A Model for Cumulative Line Delay in Full Recovery Condition

Two fundamental measures in schedules are typically used to improve reliability reducing the risk of primary delays, and damping the propagation into secondary delays; *supplements* and *buffers* (Goverde and Hansen, 2013). A supplement is an additional time beyond the minimum operating time between timing points that allows a train to experience disruptions and yet still attain scheduled arrivals (Figure 3.1-1). This measure is specific (and potentially unique) to each train between a given pair of timing points and directly supports timetable robustness.

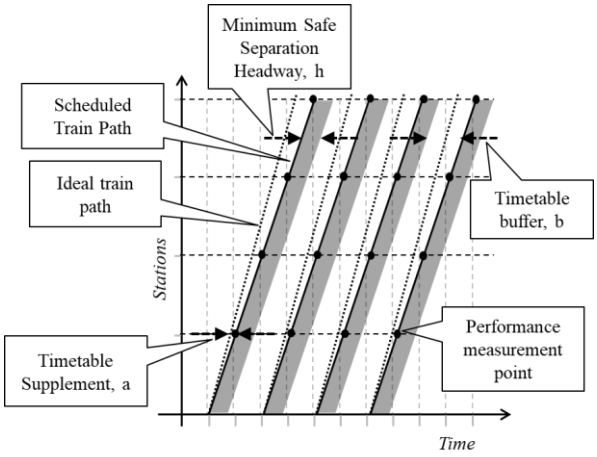

Figure 3.1-1: Definition of timetable supplement and headway buffer in train paths.

A *buffer* is an additional time between trains so that disruptions and delays of the leading train are less likely to cause interference with the following train (Figure 3.1-1). The buffer is a component of the headway (the total time between passing trains), but not the same as the headway. The headway equals the minimum safe separation time between trains plus the buffer. The capacity or number of trains on the railway line is strictly determined by the headway, but clearly, the buffer is a decision variable, that, other things being equal, determines the tradeoff between capacity and stability.

Delays to trains are classified as *primary* or *secondary*. Primary delays are events happening to or "owned" by a specific train, such as a driver mistake, a passenger incident, unusual crowds, etc. Secondary delays are delays experienced as the result of conflict or interference with another train that has deviated from its plan (Figure 3.1-2).

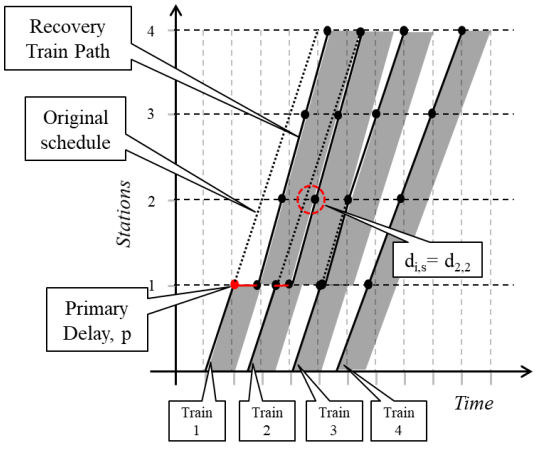

Figure 3.1-2: Calculation of secondary delay as a function of primary delay.

In the analysis that follows, the total cumulative delay of the timetable system is calculated. That is the total deviation from the timetable of every train at every station (measuring point) over the time horizon of the analysis period. Thus, a train that is five minutes late at three sequential stops would register fifteen minutes of system delay.

Additionally, the formulation provided returns the *settling time* from a primary delay. The settling time is the time it takes from a primary delay until the trains have returned to their planned schedules. The measure is used in timetable robustness assessment. For example, Salido et al. (2008) define a settling time performance measure called *(t,k) robust*. The measure states that if a timetable suffers a disruption of *t* time units and returns to stability or the original plan in *k* time units, then it is (*t*,*k*) robust.

3.1.3.1 A Finite Series Model of Delay in Two Dimensions

This section proposes a model that is a closed form function that calculates the total delay as a function of a single initial delay to one train. The model is, then, extended in section 3.1.4 to include multiple primary delays at any location on the line.

Many of the prior cited papers define the analysis horizon in terms of the length of line or the number of train path segments. The following model specifically includes the secondary delays to individual trains, and thus the second dimension of the analysis horizon is the train number in a sequence. This model consider trains on a single line with a single direction of movement (e.g. on a double-track railway), which is a conventional operating plan in Europe and urban North America, and is likewise matter of study of mentioned literature (Carey, 1999; Cerreto, 2015; Hasegawa et al., 1981; Huisman and Boucherie, 2001; Landex, 2008; Mattsson, 2007). The time horizon of the model then begins with the train and location of the primary delay and ends with the return of the last train to schedule within the allowed delay threshold.

Table 3.1-2 presents the terms and labels for sets and parameters in the model. Without loss of generality, the timetable measurement points are called “stations”, even though they can just as well be any geographic location where the train must adhere to the timetable. Subscripts *i* and *s* specify the train and station that the parameters refer to, respectively. When parameters are later used without a subscript, they are constant and identical for all trains in the formulation, so that *a*, for example, refers to a value of running time supplement constant and identical for every train between any pair of consecutive stations. s^*_i refers to the last station after a disruption at which train *i* deviates from its planned timetable, namely its last delayed station. i^*_s is a companion component of the last delayed station and refers to the last train at a given station after a disruption which

deviates from its planned timetable. δ is the delay threshold, which means that delays below a specified magnitude will be ignored in the calculation of utility loss.

s	Station index
i	Train index
S	The ordered set of stations of the analysis, {1,2,…,S}. s=1 is the station where the primary delay is generated.
I	The ordered set of trains in the analysis, {1,2,…,I}. Lower numbered trains precede higher numbered trains. i=1 is the train the receives primary delay.
p	Primary delay
$d_{i,s}$	Individual delay of train *i* at station *s*
$t_{i,s}$	Minimum running time of train *i* between stations *s-1* and *s*
$h_{i,s}$	Minimum time separation (headway) between trains *i-1* and *i* at station *s*
δ	Delay threshold
$a_{i,s}$	Running time supplement of train *i* between stations *s-1* and *s*
$b_{i,s}$	Headway buffer time at station *s* between trains *i-1* and *i*
s^*_i	Last delayed station for train *i*
i^*_s	Last delayed train at station *s*
ω	Timetable slack ratio
φ_p	Timetable settling time for delay *p*
Γ	Cumulative line delay

Table 3.1-2: Table of sets and parameters

3.1.3.2 Primary Delays and Derivation of Cumulative Delay

Primary delays can occur at any station of the line and affect any train in the schedule. A residual amount of delay that exceeds the possible recovery by running time supplement and headway buffer propagates to succeeding trains and downstream stations. This section models the relation between individual train delays to calculate the cumulative line delay. Train and station indices *i* and *s* start at 1 at the location of the primary delay.

The cumulative delay, *Γ,* represents the unweighted utility loss experienced by the railway service due to a disruption. It is the sum of all individual delays as measurement points in the timetable over the analysis horizon and is presented in Equation (1).

$$\Gamma = \sum_{\substack{i \in \boldsymbol{I} \mid d_{i,s} \geq \delta \\ s \in \boldsymbol{S}}} d_{i,s} \tag{1}$$

p defines the primary delay, corresponding to the delay of the first train at the first station in the analysis horizon, thus $d_{1,1}=p$. This delay will propagate to the following trains through the expression of individual delay given in (2)

$$\begin{gathered} d_{i,s} = max\{(d_{i,s-1} - a_{i,s}), (d_{i-1,s} - b_{i,s}), 0\} \ \forall\, i > 1, s > 1 \\ d_{1,s} = max\{(d_{i,s-1} - a_{i,s}), 0\} \ \forall\ s > 1 \\ d_{i,1} = max\{(d_{i-1,s} - b_{i,s}), 0\} \ \forall\, i > 1 \\ d_{1,1} = p \end{gathered} \tag{2}$$

Equation (2) represents that every train delay incidence is caused either by delay originating with the train or by secondary delay imposed by another train. A train may recover from its own delay by the timetable supplement $a_{i,s}$. A train is likewise shielded from the obstruction of a preceding train by the buffer $b_{i,s}$. A train then experiences a delay if either or both of these limits are exceeded, and the larger of the two values determines the functional train delay. Similarly to previous research (Goverde, 2010, 2007; Hasegawa et al., 1981; Huisman and Boucherie, 2001; Landex, 2008; Mattsson, 2007), delay recovery is modeled here as a deterministic process. Nevertheless, the stochasticity of delay recovery can be included in the model using expected values of delay recovery in place of scheduled running time supplement and headway buffer. The values of expected delay recovery might be extracted from historical data.

3.1.3.3 Relaxed Formulation for Homogeneous Train Schedules and Line Segments

Homogeneous timetables are characterized by a repetition of identical train trajectories equally distributed over time. This type of schedule is very frequent in specialized railway lines, where the service pattern is constant, such as regional and suburban railway lines, metro services, or even dedicated high-speed railway lines. In these cases, the running and supplement and headway buffer are equal for all the trains and can be generalized in $a_{i,s} = a_s$ and $b_{i,s} = b_s$.

Consider now a theoretical railway line where running time supplements and headway buffers are identical throughout all the line segments and stations, such that a and b are constant values throughout the formulation. The assumption does not have an effect on minimum and scheduled running times, nor does it affect minimum and scheduled headway, which can vary freely if the timetable slack is kept constant throughout the line. Heterogeneous timetables, where running time supplements and

headway buffers vary across pairs of trains and stations can be reduced to pseudo-homogeneous using aggregate measures of the two types of slack. Landex (2008) uses the arithmetic mean of the buffer time as an aggregate measure of timetable slack in his delay propagation model. In this paper, the weighted averages of running time supplements and headway buffers are proposed instead, with weights inversely proportional to the distance from the primary delay location. The case study in section 3.1.5 shows that such aggregation is more accurate than the arithmetic mean proposed by Landex. The system-wide parameters *a* and *b* could also derive from statistical analysis of historical data from the real operation. Delay recovery would be thus accounted for as influenced by stochastic input, and the net parameters would be the expected valued of delay recovery. Equation (2) becomes Equation (3)

$$d_{i,s} = p - (s-1)a - (i-1)b \;\forall\; i \geq 1, s \geq 1 \mid p \geq (s-1)a + (i-1)b$$
$$d_{i,s} = 0 \; elsewhere \tag{3}$$

The conditions linking *p*, *i*, *s*, *b*, and *a* define a two-dimensional region where individual train delay, $d_{i,s}$, is positive. Outside this region, trains have returned to their original planned timetable, or recovered. This region, where trains are recovering or settling back into their planned timetable, is defined as the *recovery region*.

Consider that a positive value of δ further defines a recovery region only where $d_{i,s} \geq \delta$, and Equations (4) and (5) yield solutions of Equation (3) for the boundary values of the recovery region in dimensions of the number of stations and the number of trains. The extreme points of the recovery region are defined at s^*_1 and i^*_1. Since both *i* and *s* must be integers, these solutions are returned as floor functions.

$$s_i^* = \left\lfloor \frac{p + b - \delta}{a} - i\frac{b}{a} \right\rfloor + 1 \mid p \geq a + \delta \tag{4}$$

$$i_s^* = \left\lfloor \frac{p + a - \delta}{b} - s\frac{a}{b} \right\rfloor + 1 \mid p \geq b + \delta \tag{5}$$

Figure 3.1-3 depicts the approximate boundaries of the recovery region. The diagonal boundary has an approximate slope of *a/b*.

One additional simplifying hypothesis is necessary to calculate the settling time. The line consists of equitemporal (not necessarily equidistant) line segments and identical train dynamic performance, such that *h* and *t* are also constant values throughout the formulation. Such an assumption is rather akin to previous hydrodynamic models applied to high-speed networks with homogeneous traffic (Hasegawa et al., 1981). The

approximate settling time will be the greater of the times necessary to traverse and exit the recovery region, either along the station axis (*s*) or the train axis (*i*), equation (6).

$$\varphi_p = \max\{(t+a)(s_1^*+1), (h+b)(i_1^*+1)\} = \max[(a+t)(2+\frac{p-\delta}{a}), (b+h)(2+\frac{p-\delta}{b})] \quad (6)$$

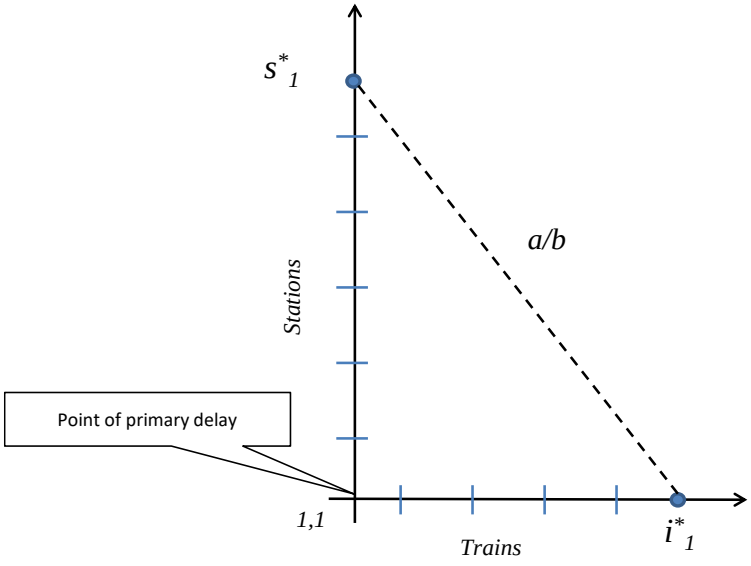

Figure 3.1-3: Recovery region bounds for a given primary delay p.

3.1.3.4 Fixed ratio b/a and Symmetric Systems

For illustrative purpose, hypothetical systems are here introduced with a fixed ratio $\omega = \frac{b}{a}$. In such systems, delay propagation equations are simplified, and dimensionality is reduced, so inference on the polynomial model be shown on a bi-dimensional graph $[p, \Gamma(p)]$.

Consider the system where $\omega = 1$, so *a*=*b*, which will be called "symmetric", because not only are the control values of timetable supplement and buffer equal, but the ratio *b/a* implies the region is symmetric with respect to the number of trains and stations. Then the recovery region is defined only by the primary delay and the single parameter *a*, as shown in Equations (7) and (8). Checking the values of s^*_1 and i^*_1, it can be seen that indeed, for $\omega = 1$, the region is symmetric with an equal number of trains and stations.

$$s^*_{i\,\omega} = \left\lfloor \frac{p-\delta}{a} + (1-i)\omega \right\rfloor + 1 \mid p \geq a + \delta \quad (7)$$

$$i^*_{s\,\omega} = \left\lfloor \frac{p-\delta}{\omega a} + (1-s)\frac{i}{\omega} \right\rfloor + 1 \mid p \geq \omega a + \delta \quad (8)$$

3.1.3.5 Calculation of Cumulative Delay

The resulting summations for the cumulative delay, Γ, are shown for the general case in Equation (9), and for the case of fixed ratio *b/a* in Equation (10). The summation operates first in the dimension of the stations for individual trains, returning the cumulative delay recorded on one train's whole itinerary, and then it sums the cumulative delay across the individual train itineraries.

$$\Gamma = \sum_{\substack{i\in\left\{1,2,\ldots,\left\lfloor\frac{p+a-\delta}{b}-s\frac{a}{b}\right\rfloor+1\right\} \\ s\in\left\{1,2,\ldots\left\lfloor\frac{p-\delta}{a}\right\rfloor+1\right\}}} p+a+b-sa-ib \tag{9}$$

$$\Gamma_{\omega} = \sum_{\substack{i\in\left\{1,2,\ldots,\left\lfloor\frac{p-\delta}{\omega a}+\frac{(1-s)}{\omega}\right\rfloor+1\right\} \\ s\in\left\{1,2,\ldots\left\lfloor\frac{p-\delta}{a}\right\rfloor+1\right\}}} p-a(\omega(i-1)+s-1) \tag{10}$$

The floor functions in these summations prevent them from resolving into manageable functions. If the floor functions are relaxed, the summations resolve into the following polynomials: the general case in Equation (11), and the case of fixed ratio *b/a* in Equation (12).

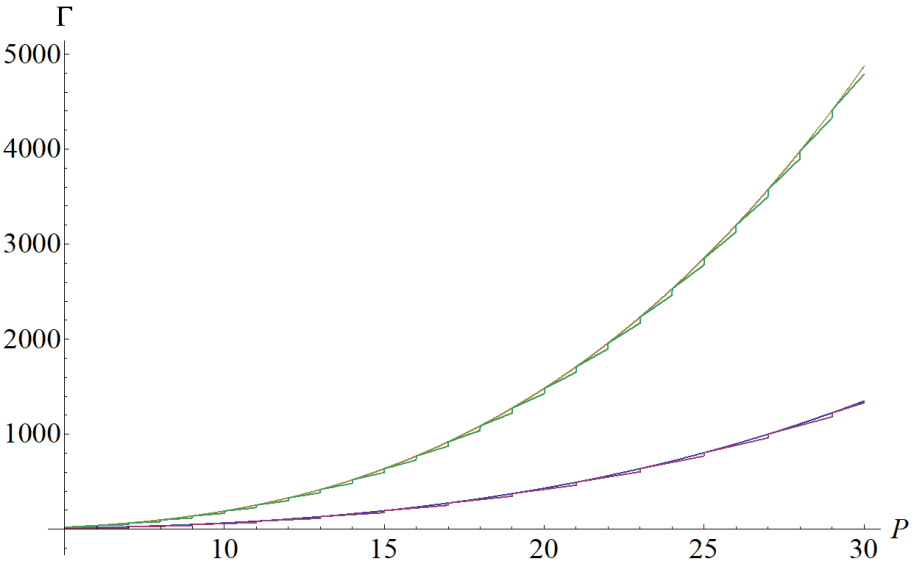

Figure 3.1-4: Plot of cumulative delay for δ=3 and fixed ratio b/a = ω =1. The fixed ratio model is used for readability. Comparison between the polynomial model and the numerical summation, with a=1 (top curve) and a=2 (bottom curve).

$$\Gamma = \frac{p^3}{6ab} + \frac{3(a+b)p^2}{12ab} + \frac{(a^2+3ab+6b\delta-6\delta^2)p}{12ab} + \frac{-a^2\delta+9ab\delta-3a\delta^2-9b\delta^2+4\delta^3}{12ab} \quad (11)$$

$$\Gamma_\omega = \frac{p^3}{6a^2\omega} + \frac{3a(1+\omega)p^2}{12a^2\omega} + \frac{(a^2+3a^2\omega+6a\omega\delta-6\delta^2)p}{12a^2\omega} + \frac{-a^2\delta+9a^2\omega\delta-3a\delta^2-9a\omega\delta^2+4\delta^3}{12a^2\omega} \quad (12)$$

Naturally, there is a question of how much error is introduced by relaxing the floor functions. Figure 3.1-4 shows that the difference, with and without the floor function, is very small for a delay threshold of 3 (minutes) and supplements and buffers of one or two (minutes).

3.1.3.6 Inferences from the Polynomial Function

Figure 3.1-5 presents the contour of a system with fixed ratio $\omega=1$ and shows that while timetable slack certainly is valuable in damping the damage of primary delays, its incremental value quickly declines. The figure suggests that supplements and buffers can be applied excessively, wasting resources without accomplishing proportional reductions in system delay.

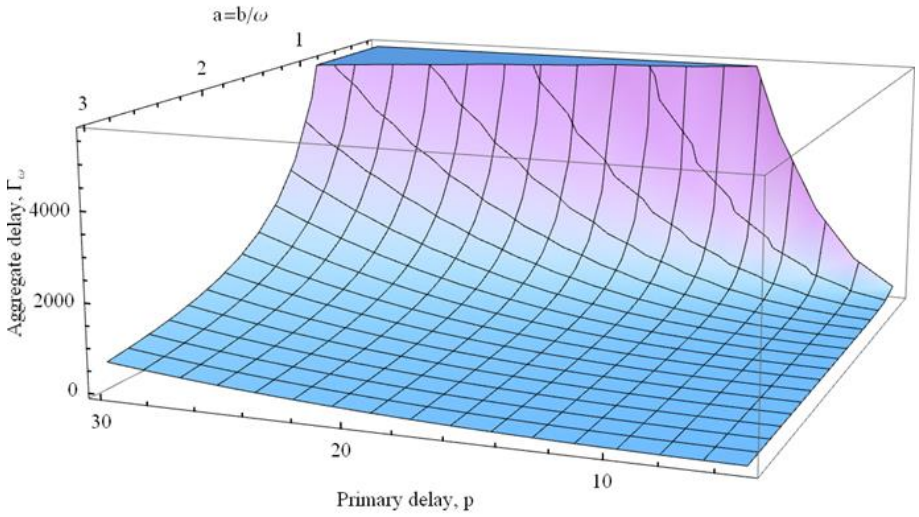

Figure 3.1-5: Aggregate delay contour of the case for $\omega=1$ and $\delta=3$, as a function of primary delay, p, and slack parameter, $a=b/\omega$.

Differential calculus of aggregate delay against timetable slack expresses the marginal reduction of aggregate delay given by increments of timetable slack. Equations (13) and (14) show the partial derivatives of cumulative against *a* and *b*, respectively. The case with fixed slack ratio α is represented by equation (15) and depicted in Figure 3.1-6.

$$\frac{\partial \Gamma}{\partial \mathrm{a}} = \frac{-\frac{p^3}{6b} - \frac{p^2}{4} - \frac{p\delta}{2} + \frac{3\delta^2}{4} + \frac{p\delta^2}{2b} - \frac{\delta^3}{3b}}{a^2} + \frac{p}{12b} - \frac{\delta}{12b} \tag{13}$$

$$\frac{\partial \Gamma}{\partial \mathrm{b}} = \frac{-\frac{p^3}{6a} - \frac{p^2}{4} - \frac{ap}{12} + \frac{a\delta}{12} + \frac{\delta^2}{4} + \frac{p\delta^2}{2a} - \frac{\delta^3}{3a}}{b^2} + \frac{-\frac{p^2}{4a} - \frac{p}{4} - \frac{p\delta}{2a} - \frac{3\delta}{4} + \frac{3\delta^2}{4a} + \frac{3ap + 3p^2 + 9a\delta + 6p\delta - 9\delta^2}{12a}}{b} \tag{14}$$

$$\frac{\partial \Gamma_\omega}{\partial \mathrm{a}} = \frac{-\frac{p^2}{4} - \frac{p^2}{4\omega} - \frac{p\delta}{2} + \frac{3\delta^2}{4} + \frac{\delta^2}{4\omega}}{a^2} + \frac{-\frac{p^3}{3\omega} + \frac{p\delta^2}{\omega} - \frac{2\delta^3}{3\omega}}{a^3} \tag{15}$$

Equations (13) and (14) show that the partial derivatives of aggregate line delay tend asymptotically to infinite for *a=0* or *b=0*, respectively. The relation is inverse of the second degree, and the hyperbolic relation suggests that the damping effect of delay propagation decreases quickly in the low range of timetable slack. The theoretical global minimum reduction of aggregate delay corresponds at infinite values of *a* and *b*, although a near-zero plateau is typically reached after the initial drop. In the model, the asymptote at null slack is due to infinite delay propagation given by either null running time supplement or headway buffer. In the first case, trains cannot recover from delays along their journey. In the second case, the delay would transfer to an infinite number of following trains. Figure 3.1-6 shows that excessive timetable slack does not contribute to reducing the aggregate line delay, reaching a plateau of near-zero marginal decrement. The plateau corresponds to timetable slack large enough to prevent any delay propagation to following trains or downstream stations. In such a system, the only delay registered is the primary delay assigned to the first train at the first station, which is unavoidable by means of timetable slack.

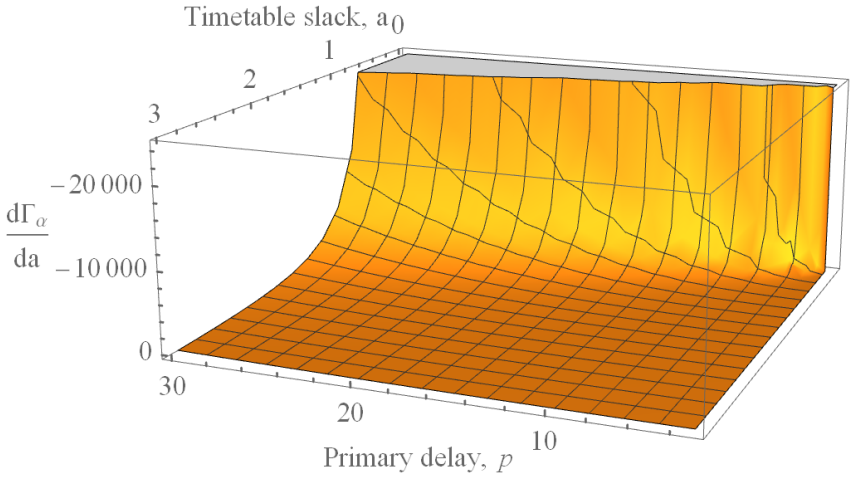

Figure 3.1-6: Contour of marginal reduction of aggregate delay against timetable slack and primary delay, with δ=3.

Further evidence that too large timetable slack is not advisable is given by the settling time formulation in equation (6). The settling time results from the maximum between two terms, which exclusively depend on the running time supplement *a*, and the headway buffer *b*, in turn. In both the terms, the relationship with the respective form of slack results from the sum of a linear term and a hyperbolic function. This suggests that, after an initial quick drop of the settling time, increases of timetable slack imply an increase in the settling time as well. An example is offered in Figure 3.1-7.

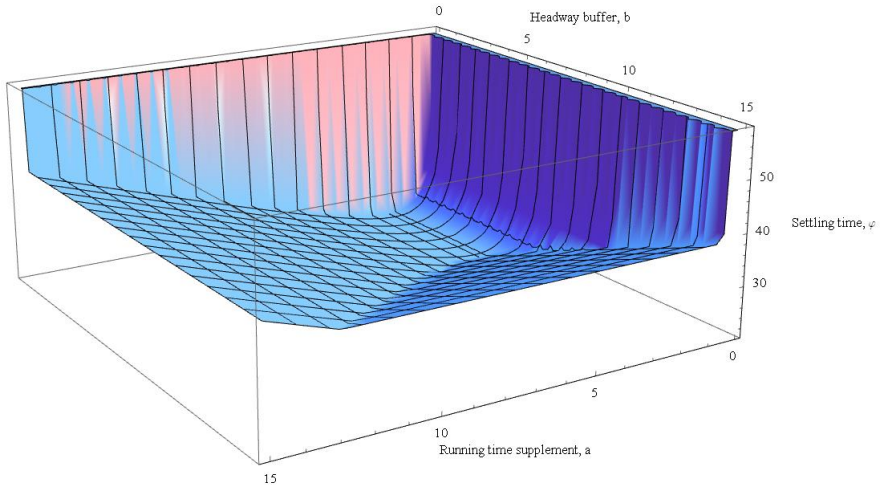

Figure 3.1-7: Contour of settling time with t=5, h=3, δ=3, and p=5.

This contour plots Equation (6) with a minimum running time between stations of *t=5*, a minimum headway of *h=3*, and a delay threshold of *δ=3*, for a given primary

delay of *p=5*. Note that both timetable supplement (a) and buffer (b) must be present to control the settling time and that excessive values of either actually worsen the settling time.

3.1.4 A Universal Polynomial Form for Primary Delays at Unspecified Stations (Any Recovery Condition)

The polynomial model presented in section 3.1.3.1 applies to full recovery conditions, i.e. all the train recover completely from delay within the study region. In real operation, incidents occur at different locations on a railway line, and trains can experience primary delays at any station. In specific cases, the delay cannot be recovered within the study region, which will be referred to as "partial recovery" condition. In this section, a universal equation is derived as an expansion of the previous case, to reduce the summation domain to a restricted study region. The equation is valid in any recovery condition and can be used to analyze the effects of primary delays at different locations on a railway line or selecting specific areas of interest. The delay recovery region is split into sub-regions, and the polynomial form is integrated with logical functions to include or exclude specific sections from the delay summation domain.

The fundamental formulation of aggregate delay keeps the same form as equation (9), which is applied to different summation domains. In the following sections, the study region and recovery regions are described, and the cumulative delay is calculated over selected portions of the domain of trains and stations modifying the summation limits. Moreover, the model parameters can be modified to consider different recovery regions originated by multiple primary delays.

3.1.4.1 Study region and Delay recovery region

Delay propagation studies can be limited to defined sections of the railway lines. For example, the most congested section of railway lines within a node could result in greater interest than marginal lines. In other cases, the lines can be divided into different homogeneous study regions, suburban railway networks can be split into sections according to the scheduled traffic volume and average headway between trains. The study region is the domain of interest in the two dimensions of the model, stations and trains, and is defined by the number of stations *S* and the number of trains *R* considered. The recovery region, defined in section 3.1.3.3, is the set of trains *i* and stations *s* where the individual delay exceeds a given threshold δ. The recovery region shapes as a pseudo-triangle in the *(i,s)* domain, which vertices are the first train at the first station, where the primary delay is generated $(1,1)$, the last delayed train at first station $(i_1^*, 1)$, and the last

delayed station for the first train $(s_1^*, 1)$. The cumulative delay is the summation of the individual delays of trains at stations within the study region, so the summation domain extends to the area resulting from the overlap of the study region and the recovery region. The overlap of study region and the recovery region depends on the values of primary delay *p*, running time supplement *a*, headway buffer *b*, delay threshold δ, number of stations *S* and number of trains *I* in the study region. Keeping the definition of cumulative delay as the overall effect of a primary delay over the study region, the equations proposed in section 3.1.3.1 are modified to include an unbounded study region where full recovery is always possible. The result is defined unbounded cumulative delay as the recovery region can extend limitless. To reduce the cumulative delay and only include the study region, individual sub-recovery areas are identified, and removed from the unbounded cumulative delay. The individual sub-recovery regions are removed when the system of equations meats specific requirements, described in section 3.1.4.3.

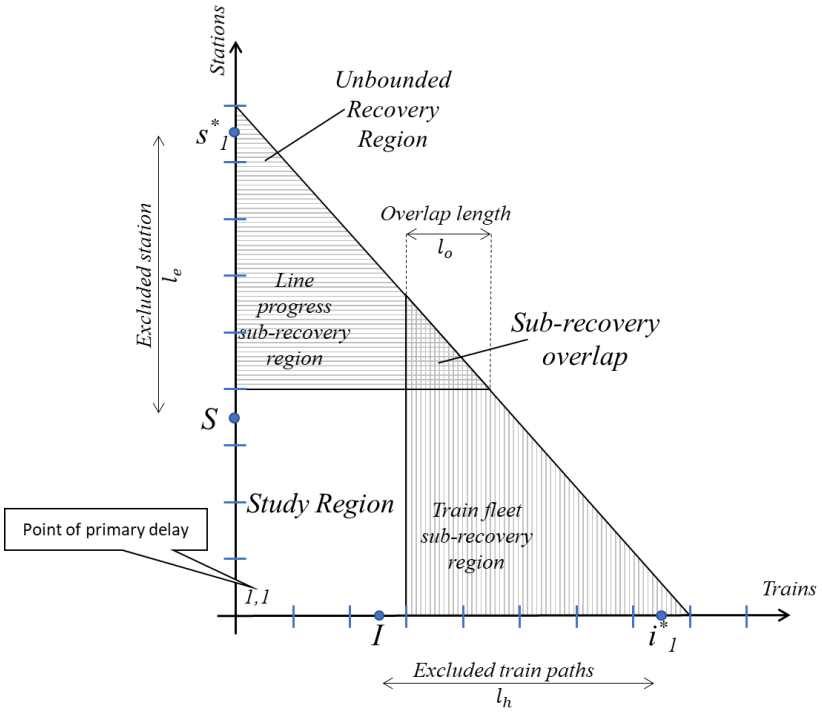

Figure 3.1-8: Study region and recovery region overlap

Figure 3.1-8 depicts a study region entirely included in the recovery region. This case shows all the possible delay sub-recovery areas to include in the general formulation for the cumulative delay. The individual areas and the specific equations are explained in

the following sections. The general formulation of the cumulative delay over individual areas keeps the same form of (9), as a summation of individual delays, whereas the summation domains differ across sub-regions.

A parametric notation is introduced here to calculate the cumulative delay over the different selections of the *(i,s)* domain. In the following sections, the expression $\Gamma_{(p,[i',i''],[s',s''])}$ indicates the cumulative delay Γ given by primary delay p at location *(i',s')*, with summation domains $i \in \{i', \dots, i''\}$ and $s \in \{s', \dots, s''\}$.

3.1.4.2 Unbounded cumulative delay and sub-recovery regions

If no restriction is imposed on the number of trains and stations, the delay will always be completely recovered. The formulation of unbounded cumulative delay corresponds to equation (9), where the summation domain extends from the point of primary delay to the last delayed train and station. The relation between primary delay and cumulative delay is, so, third degree. The unbounded cumulative delay is defined in parametric notation as $\Gamma_{(p,[1,i_S^*],[1,S])}$.

When the study region terminates before the last delayed station, namely $s_1^* > S$, the unbounded cumulative delay must be reduced by an amount corresponding to the cumulative delay in the exceeding area, called here Line progress sub-recovery. The summation domain extends in this case from the first station outside the study region $S + 1$ to the last delayed station for the first train s_1^*. Equation (16) shows the parametric notation for the Line progress sub-recovery region and the resulting polynomial.

$$\begin{aligned}\Gamma_{(p,[1,i_S^*],[S+1,s_1^*])} &= \sum_{\substack{i\in\{1,\dots,i_S^*\}\\ s\in\{S+1,\dots,s_1^*\}}} p + a + b - sa - ib \\ &= \frac{p^3}{6ab} + \frac{p^2(3a+3b-6aS)}{12ab} \\ &\quad + \frac{p(a^2+3ab-6a^2S-6abS+6a^2S^2+6b\delta-6\delta^2)}{12ab} \\ &\quad + \left(\frac{-a^3S-3a^2bS+3a^3S^2+3a^2bS^2-2a^3S^3-a^2\delta}{12ab}\right. \\ &\quad \left. + \frac{9ab\delta-6abS\delta-3a\delta^2-9b\delta^2+6aS\delta^2+4\delta^3}{12ab}\right)\end{aligned} \tag{16}$$

Similarly to the line progress sub-recovery, when the study region terminates before the Last Delayed Train, namely $i_1^* > R$, the unbounded cumulative delay must be reduced by an amount corresponding to the cumulative delay in the exceeding area, called

here Train fleet sub-recovery. The cumulative delay in the Train fleet sub-recovery region is denoted as $\Gamma_{(p,[R+1,i_S^*],[1,s_1^*])}$.

In the general case, the mentioned exceeding areas can overlap. This happens when the delay cannot be recovered by any of the trains in the study region, before the last station. In these cases, the overlapping excess should be reintroduced to avoid double subtraction. The last delayed station for the first train outside the study region is a new type of boundary for the summation domain. The formulation is derived from (4) and it is defined as s_o^* in equation (17).

$$s_o^* = s_{I+1}^* = \left\lfloor \frac{p+b-\delta}{a} - (I+1)\frac{b}{a} \right\rfloor + 1 \mid p \geq a + \delta \tag{17}$$

The summation domain extends from the first station outside the study region to s_o^*, and the parametric notation of the related cumulative delay is $\Gamma_{(p,[I+1,i_S^*],[S+1,s_o^*])}$.

3.1.4.3 Universal formulation

The existence of the individual sub-recovery areas mentioned in this section depends on the relation between system parameters, *I*, *S*, *a*, *b*, δ, and the primary delay *p*. The universal formulation proposed in this section includes logical controls on the specific delay sub-recovery to select only the regions that are active. The logical controls include the formulation from individual areas only if their specific dimension is positive. The sub-recovery regions respective control lengths are defined hereunder.

The line progress sub-recovery region is controlled by l_e, the number of excluded stations, defined in (18) as the difference between the last station in the study region and the last delayed station for the first train. Similarly, the line progress sub-recovery region is controlled by l_h, the number of excluded train paths, defined in (19) as the difference between the last train in the study region and the last delayed train at the first station. The sub-recovery overlap is controlled by l_o, the overlap length between the last delayed station for the first train outside the study region and the last station in the study region. The overlap length is calculated by equation (20).

$$l_e = s_1^* - S = 1 - S + \frac{p-\delta}{a} \tag{18}$$

$$l_h = i_1^* - I = 1 - I + \frac{p-\delta}{ab} \tag{19}$$

$$l_o = s_o^* - S = 1 - S + \frac{b+p-b(1+R)-\delta}{a} \tag{20}$$

Equation (21) is a closed form function that returns the cumulative delay on a railway line as a function of the primary delay *d(p)*, in any condition of recovery, given the system variables *a*, *b*, *δ*, *R*, *S*.

$$\Gamma_{(p,[1,R],[1,S])} = \Gamma_{(p,[1,i_S^*],[1,S])} * \frac{\max(s_1^*, 0)}{s_1^*} - \Gamma_{(p,[1,i_S^*],[S+1,s_1^*])} * \frac{\max(l_e, 0)}{l_e} - \Gamma_{(p,[R+1,i_S^*],[1,s_1^*])} * \frac{\max(l_h, 0)}{l_h} + \Gamma_{(p,[R+1,i_S^*],[S+1,s_O^*])} * \frac{\max(l_o, 0)}{l_o} \quad (21)$$

The logical control of the unbounded cumulative delay is introduced in this formulation as a replacement of the constraint $p \geq (s-1)a + (i-1)b$ in (3), to simplify the formulation.

3.1.4.4 Model application on heterogeneous lines and networks

The analytical model presented in this paper is formulated for homogeneous railway lines, where timetable parameters *a* and *b* have similar values across trains and line sections and can be summarized in unique input values for equation (21). However, the model applicability is not limited to the mentioned case and can be extended to other types of railway networks, after partitioning the networks into homogeneous sections to be modeled recursively.

In this section, the conditions that rule the model application on heterogeneous lines are analyzed. A hypothetical railway line is divided into two homogeneous sections, with specific traffic volume, number of stations, and timetable parameters *a* and *b*. The sections are named *A* and *B*, and the set of stations $\boldsymbol{S}$ is divided consequently in $\boldsymbol{S_A} = \{1,2,\dots,S_A\}$ and $\boldsymbol{S_B} = \{S_A+1,\dots,S\}$. The traffic volume, running time supplement and headway buffer are defined independently for the two sections by I_A, I_B, a_A, a_B, b_A, and b_B, respectively. A new timetable parameter is introduced, defining the traffic volume ratio between the two line-sections, $\rho = \frac{I_B}{I_A}$. Note that when $\rho<1$ the traffic volume is smaller in the second section, meaning that all the services run in the first section with every *(1/ρ)-th* train continuing in the second section. When $\rho>1$ all the trains end at the same destination, I_A of which originate from the first station, and $(\rho-1)* I_A$ originate at the second section, equally distributed in the timetable. Timetable discontinuities can also manifest when the traffic volume does not change, so $\rho=1$, if, for example, either or both *a* and *b* are remarkably different across to line-sections.

Figure 3.1-9 illustrates the two traffic conditions with $\rho>1$ and $\rho<1$.

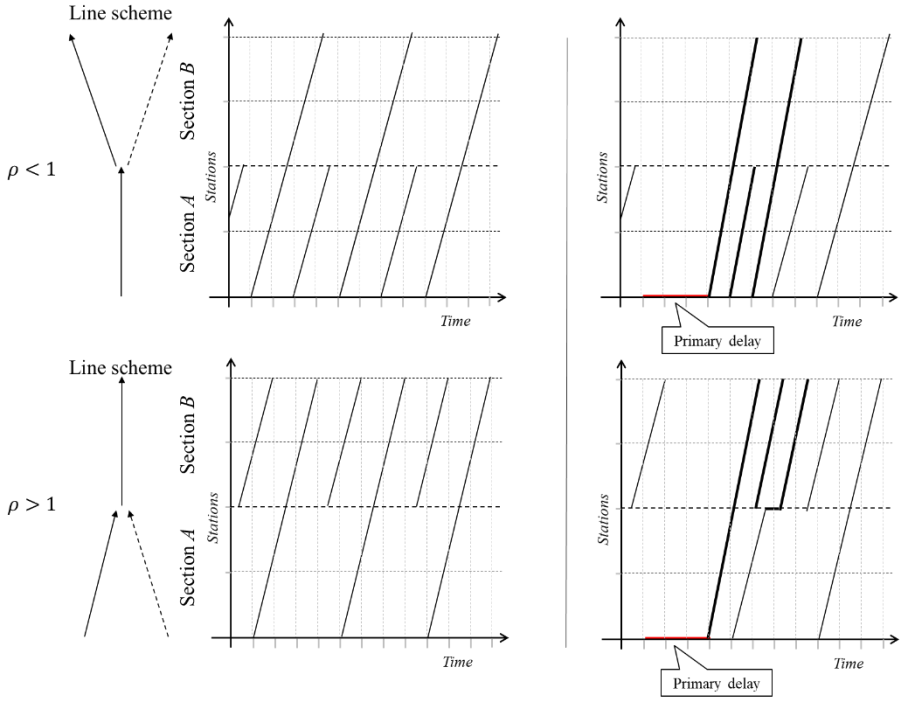

Figure 3.1-9: Model application on lines with heterogeneous traffic. Upper row represents diverging lines with lower traffic in section B. Lower row represents converging lines with greater traffic volume in section B. Left side: unperturbed traffic; Right side: primary delay in red, delayed trains bold. Time-space diagram on dashed branches omitted for readability.

Individual train delays and cumulative line delay in section *A* are determined by equations (3) and (21) replacing *S* with S_A. Delay propagation across sections *A* and *B* depends on the relations between traffic volumes and headway buffers between the two sections. The hypothesis of equal headway buffer is not valid in the discontinuity between stations S_A and $s_{(A+1)}$ so the individual delay propagation is ruled by equation (2). At the first station in section *B*, the residual delay of trains from originated in section *A* is exceeded by the knock-on delay from previous trains under specific conditions:

- When $b_A > \rho * b_B$, the knock-on delay from previous trains is greater than residual delay from section *A*, and the delay propagation in section *B* is only ruled by the residual delay of the first train. A single application of equations (3) and (21) describes the entire traffic in section *B*.
- When $b_A < \rho * b_B$, the residual delay from section *A* is greater than the knock-on delay from previous trains, and the delay propagation in section *B* is overrun by the residual delay of every train originated in section *A*. A

separate delay propagation model is necessary for section *B* starting at every train originated in section *A*.

Note that for $\rho>1$ the assumption of unmodified train order is necessary in case of delays. In real operation, the assumption is reasonable with small delays, where alterations of the train sequences would cause larger disruption than the delay itself. Larger primary delays might be modeled with the support of queuing theory from other models and introducing dispatching criteria to select the train that passes first in case of conflicts of convergent itineraries.

The process described in this section extends the model applicability. Diverging lines can be considered as sets of homogeneous subsections, where the traffic volume is larger in the first section, and the model can be applied separately in the diverging section. Converging lines can be modeled as cases where $\rho>1$ and require the same assumption of unchanged train sequence.

Similarly, discontinuities in the timetable structure may be addressed considering a separation of the study region into homogeneous sub-regions in the dimension of trains $\boldsymbol{I \rightarrow I_A, I_B}$.

3.1.4.5 Multiple primary delays

The universal formulation of cumulative line delay in equation (21) provides the model the flexibility to consider study regions of unspecified dimensions in the domain of trains and stations. One of the advantages is the adaptability of the model to different recovery conditions, and the possibility to investigate the effects of individual primary delays on selected portions of the study regions. This section presents methods to consider multiple simultaneous delays, based on individual calculations of cumulative line delay on portions of the study region. Two different cases are identified, according to the relative position of primary delays in the domain of trains and stations.

Consider two primary delays, p_1 and p_2, generated by independent events taking place at stations s_{p1} and s_{p2}, on trains i_{p1} and i_{p2}, respectively. Defined the distances between the primary delay events in the dimension of stations and trains, respectively $\Delta s = s_{p2} - s_{p1}$ and $\Delta i = i_{p2} - i_{p1}$, the ratio $\frac{\Delta s}{\Delta i}$ determines the simultaneity case, considering that perturbations only propagate to downstream stations and to successive trains. Figure 3.1-10 depicts the two cases and the related metrics.

For $\frac{\Delta s}{\Delta i} > 0$, the effect of p_2 falls entirely in the propagation area of p_1 and will be named therefore nested simultaneous primary delay. p_2 is considered as a further delay

over existing residual delays from previous perturbations. The effects of the two primary delays cannot be summed, due to

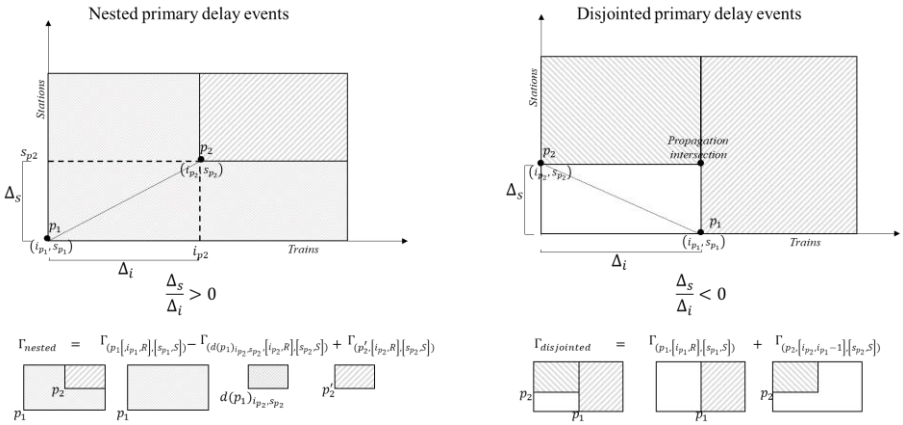

Figure 3.1-10: Relative position of simultaneous primary delays and study region portions.

the model non-linearity. The cumulative line delay in case of nested simultaneous primary delays is given by equation (22)

$$\Gamma_{nested} = \Gamma_{(p_1,[i_{p_1},I],[s_{p_1},S])} - \Gamma_{(d(p_1)_{i_{p_2},s_{p_2}},[i_{p_2},I],[s_{p_2},S])} + \Gamma_{(p_2',[i_{p_2},I],[s_{p_2},S])} \quad (22)$$

where $\Gamma_{(p_1,[i_{p_1},I],[s_{p_1},S])}$ is the unconditioned cumulative line delay generated by primary delay p_1, $d(p_1)_{i_{p_2},s_{p_2}}$ is the individual train delay generated by p_1 on the train affected also by p_2, $\Gamma_{(d(p_1)_{i_{p_2},s_{p_2}},[i_{p_2},I],[s_{p_2},S])}$ is the effect of p_1 that must be removed and replaced by the effect of p_2. $\Gamma_{(d(p_1)_{i_{p_2},s_{p_2}},[i_{p_2},I],[s_{p_2},S])}$ is calculated through equation (21), with coordinates of primary delay i_{p2} and s_{p2}, and amount of primary delay equal to $d(p_1)_{i_{p2},s_{p2}}$. $\Gamma_{(p_2',[i_{p_2},I],[s_{p_2},S])}$ is the effect of the nested primary delay p_2, cumulated with a previous residual delay from p_1. $\Gamma_{(p_2',[i_{p_2},I],[s_{p_2},S])}$ is calculated through equation (21), with coordinates of primary delay s_{p2} and i_{p2}, and amount of primary delay equal to $p_2' = d(p_1)_{i_{p2},s_{p2}} + p_2$. Alternative approaches are possible, in the case recorded delays are available in place of incremental primary delays. In these cases, p_2' is set equal to the highest between recorded deviation from schedule and residual delay from previous primary delay.

For $\frac{\Delta s}{\Delta i} < 0$, both primary delay events propagate to a portion of the study region as unique primary events. In the propagation overlap region, the highest residual delay is counted in the cumulative line delay, considering smaller delays surpassed by larger delays after equation (2). This case will be named disjointed simultaneous primary delay, and the

study region is divided into two sub-regions. The study-area partitioning is defined by the comparison of residual delay at the intersection point in the trains-stations domain from the individual primary delays considered.

Without loss of generality, the case of greatest residual delay from p_1, is here described, assuming that $i_{p2} > i_{p1}$ and $s_{p2} < s_{p1}$. In this case, the aggregate line delay is expressed by equation (23)

$$\Gamma_{disjointed} = \Gamma_{(p_1,[i_{p_1},I],[s_{p_1},S])} + \Gamma_{(p_2,[i_{p_2},i_{p_1}-1],[s_{p_2},S])} \quad (23)$$

where $\Gamma_{(p_1,[i_{p_1},I],[s_{p_1},S])}$ is the cumulative delay generated by primary delay p_1, calculated from delay location and train through the rest of the study region, and $\Gamma_{(p_2,[i_{p_2},i_{p_1}-1],[s_{p_2},S])}$ is the cumulative delay generated by primary delay p_2, calculated from the location of delay and to the location of intersection with the residual delay from p_1, intersection excluded.

3.1.4.6 Inference on the delay threshold from the universal polynomial form

The closed form introduced in section 3.1.4 provides the model flexibility to represent delays at any location of the study region and allows to infer the effect of different values of the delay threshold in a given timetable, to evaluate the most appropriate value of delay threshold δ.

Figure 3.1-11 presents the cumulative delay function and shows the effect of different delay threshold strategies. Intuitively, no cumulative delay is recorded for primary delays smaller than the delay threshold. The figure shows also the range of effectiveness of the delay threshold. The cumulative delay measure is dampened in situations of full or partial recovery.

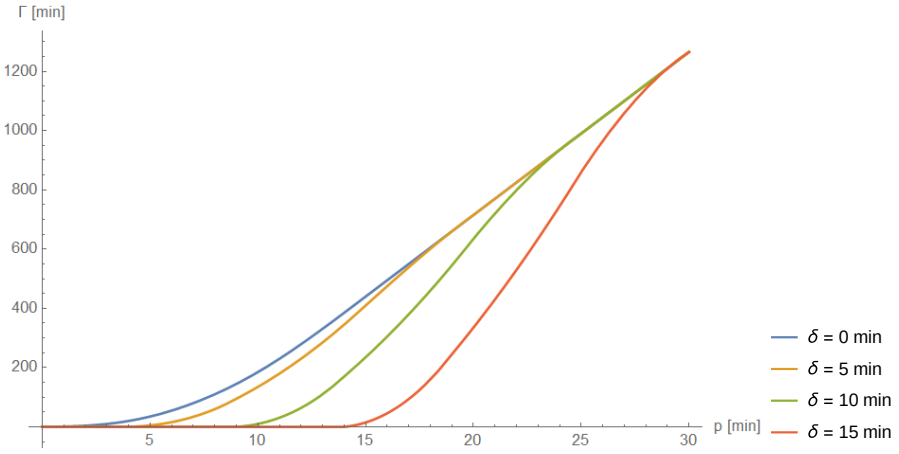

Figure 3.1-11: Cumulative delay as a function of primary delay on a railway line with different values of delay threshold. S = 11 stations, R = 5 trains, a = 1 min, b = 1 min

The delay threshold becomes ineffective as soon as all the trains are delayed enough to be included in the summation, which corresponds to meeting condition (24)

$$\begin{aligned} d_{I,S}(p) &< \delta \\ p &< \delta + a(S-1) + b(I-1) \end{aligned} \tag{24}$$

The developed closed form railway line propagation model makes it possible for service contractors and transport authorities to conveniently evaluate delay thresholds that meats the measured delay distributions and reduces operational costs improving the measured punctuality.

The delay threshold allows for some flexibility in the planning phase, under the assumption that small delays are not perceived by the passengers. The calibration of the delay threshold in service contracts between service providers and transport authorities has an influence on daily operation, and different strategies in the delay threshold dimensioning lead to different dispatching strategies to pursue the measured punctuality. Punctuality penalties are a relevant share of operations budget of transportation companies, especially in cases where a performance regime is applied. For example, the European Performance Regime (2013) draws the guidelines for performance management in the European countries, and every single minute of delay of a train can cost to the service contractor up to 2€ (Rete Ferroviaria Italiana, 2015). The relation between punctuality measurement methods, delay thresholds, and distribution of running time supplement in train paths across several countries in Europe is described by Schittenhelm (2011). Suburban and regional railway services in Europe admit thresholds between 3 and 5 minutes, whereas long distance services are allowed to reach from 5 to 15 minutes of delay before penalties are applied.

3.1.5 Case study

In this section, a contemporary suburban railway in Denmark is simulated and comparisons are made between the measured and theoretical system delay. The simulation is performed in OpenTrack (Nash and Huerlimann, 2004). The subject line is the Hillerød suburban railway on its northern segment from Hellerup to Hillerød (29 km). On this segment, there are eleven stations inclusive of the terminal, Hillerød, and the junction Hellerup. Hellerup is not the end of the line. All trains continue through Hellerup, through Copenhagen, and on to destinations much further south of Copenhagen.

In this case study, a cyclic timetable is simulated with homogeneous train traffic of all-stops services from Hellerup to Hillerød and scheduled headway of 5 minutes. Running time supplements and headway buffers are identical for every train path allocated but differ

across stations, and this is where the case study deviates significantly from the theoretical model.

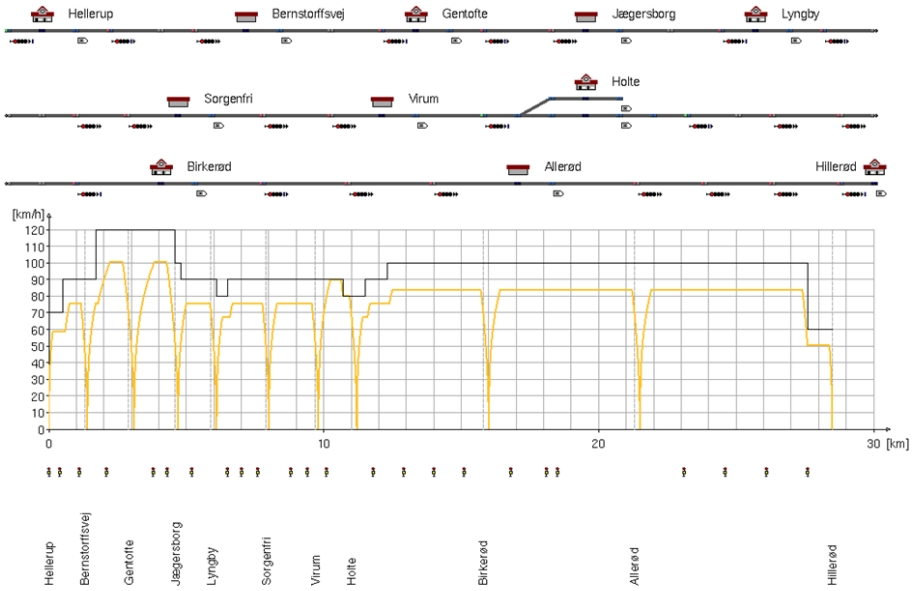

Figure 3.1-12: Simulation model of Hillerød Suburban railway in OpenTrack showing track blocks and speed profile.

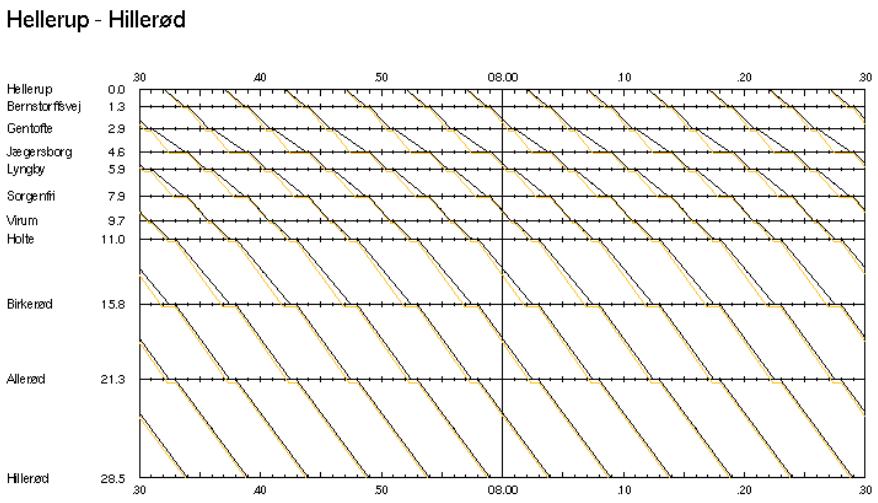

Figure 3.1-13: Graphical timetable (stringline) for Hillerød Suburban railway. Schedule in black, simulated operation colored.

Figure 3.1-12 shows the simulation model track block structure and the train performance speed profile. Figure 3.1-13 shows the graphical timetable or stringline diagram for the service between Hellerup and Hillerød.

3.1.5.1 Experimental Design of Simulation

In the simulation analysis presented, primary delays are experienced by the first scheduled service at the Hellerup station from a uniform distribution of [0,900] seconds, and 200 replications are sampled. Only northbound traffic to Hillerød is studied. The cumulative delay is measured across all the services simulated on the line. Two hours of operation are simulated, from 7:00 to 9:00, including thus 24 identical train paths.

Unlike the theoretical model of Section 3.1.3.1, the stations in this sample are not uniformly distributed. Further, the supplement and buffer times are not uniformly distributed along train trajectories, but trajectories repeat equally through time. Recalling section 3.1.3.3, homogeneous timetables are timetable where $a_{i,s} = a_s$ and $b_{i,s} = b_s$ for any value of i. Table 3.1-3 presents the actual timetable supplements and headway buffers scheduled in the simulated Hillerød Suburban railway.

Station name	**Hellerup**	**Bernstorffsvej**	**Gentofte**	**Jægersborg**	**Lyngby**	**Sorgenfri**	**Virum**	**Holte**	**Birkerød**	**Allerød**	**Hillerød**
Station number	1	2	3	4	5	6	7	8	9	10	11
Distance from origin [km]	0,0	1,3	2,9	4,6	5,9	7,9	9,7	11,0	15,8	21,3	28,5
Running time supplement from previous station a_s [s]	0	21	20	83	10	64	9	9	67	45	62
Headway buffer between trains b_s [s]	261	202	181	204	176	194	214	159	160	57	187

Table 3.1-3: Timetable supplements and headway buffers on Hillerød Suburban railway as simulated, seconds.

The closed-form model requires unique values of headway buffer and running time supplement that represent the slack structure of the timetable. The weighted averages of the buffer and supplement for all measuring point of the services is here used to aggregate $a_{i,s}$ and $b_{i,s}$ in single values a_s and b_s. Weights are assigned in inverse proportion to the distance from the location of primary delay so that the entity of slack in sections close to

the primary delay have more influence in the recovery than the further ones. In particular, weights w_s are assigned by $w_s = \frac{l_S - l_s}{l_S}$, where l_s is the physical distance between primary delay and station s, and S is the last station on the line. Resulting timetable parameters are: $a = 34{,}7\ s$ and $b = 159{,}1\ s$. The delay threshold, δ, is zero, and all delays of any magnitude are included in the cumulative delay.

3.1.5.2 Results of the Simulation

The simulation results from OpenTrack are summarized in Figure 3.1-14. Primary delays smaller than 368 s are fully recovered within the study region. Model inversion of equation (4) returns a maximum theoretical primary delay of 347 s. Larger delays cannot be recovered before the end of the line for the first train, so a line progress sub-recovery region is generated. No train fleet sub-recovery region is generated, as primary delays in the simulated range are always recovered before the last train path. Note that the model approximation is very close to the simulated result.

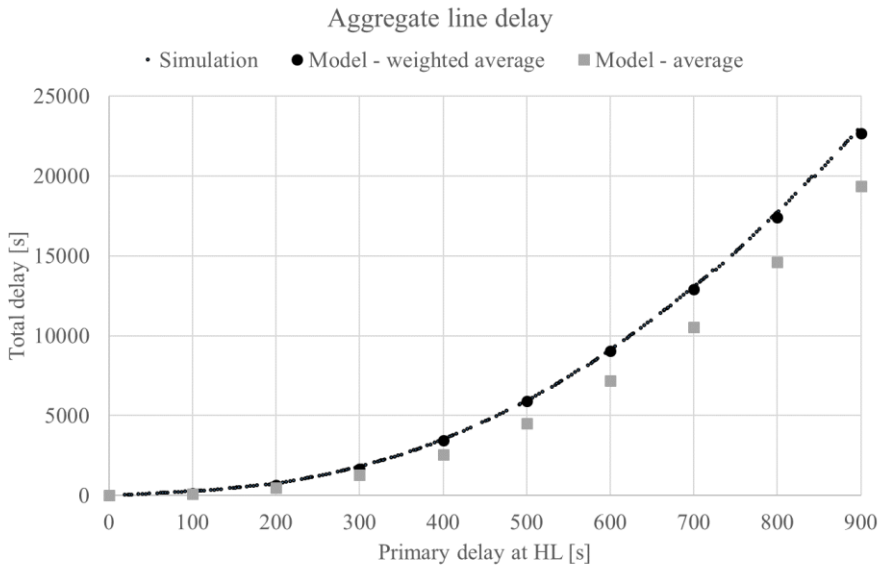

Figure 3.1-14: Comparison of the results from the simulation of delays against the estimated delay from the polynomial function.

3.1.6 Model discussion

The polynomial form proposed in section 3.1.3.1 for unbounded delay recovery is third degree. It is a cubic function of the primary delay if the measurement horizon extends fully over the recovery region. This agrees with the earlier findings of Hasegawa et al. (1981). Our paper differs from Hasegawa in that it explicitly models the discrete

summation of delays, considering three parameters: *supplement*, *buffer*, and *threshold* for measurable delay. Hasegawa's theoretical model relies on the input of spatial density, recovery flow and speed, represent timetable slack only implicitly and require simulation-based calibration of parameters. The universal model presented in section 3.1.4 results in a composite polynomial function of primary delay instead of the purely cubic function of delay in Hasegawa, only valid in case of full recovery within the study region. If the measurement horizon is restricted to less than the full recovery region, the delay excess must be removed from the recovery region, and the polynomial reduces to second degree and over very short horizons it is linear. The possibility to calculate cumulative delay over line subsections of unconstrained size lends the model the flexibility to represent traffic discontinuities on the line, simple networks, and multiple primary delays, which is not possible in Hasegawa's unbounded model.

The extended polynomial form finds a functional shape in agreement with findings from Salido et al. (2012). Salido et al. observed a linear relationship between primary delays and aggregate delays, under light utilization, which corresponds to the model presented in this paper, with study area restricted to one or a few trains. As opposed, a nonlinear relationship is found in heavily utilized lines, which corresponds to the third-degree polynomial resulting from a large number of trains involved.

The polynomial approximates the discrete summation and is robust over a wide range of parameters. Investigation of the derivatives of the polynomial finds that excessive values of running time supplement and headway buffer are ineffective in damping delay propagation and may result in poor timetable stability. This is in agreement with Carey's stochastic station delay model (1999), according to which the effectiveness of timetable slack fades out when the slack is too large, based on typical down-sloping delay distributions. A simplified formulation of the cumulative delay function is provided for fixed ratio a/b, where the only parameter is the general timetable slack. The simplified formulation is particularly valuable for illustrative purposes.

The newly introduced timetable parameters, the aggregated running time supplement a and headway buffer b can be interpreted as different measures of timetable slack or tuning parameters in the timetable to rule how the delay is recovered in the dimensions of station and trains. The model proposed in this paper supports the investigation of the effects of schedule adjustments, and the related structural changes in the delay recovery. In the case study, a weighted average of individual slack elements is proposed, based on the relative distance from the primary delay. The recovery elements

closest to the primary delay location were successfully assumed more relevant than the furthest ones, but more sophisticated methods to aggregate these parameters might be investigated in further research.

The universal form presented in section 3.1.4.4 describes the more general case where primary delays occur on unspecified trains at generic stations and allows a more extended analysis of line delays, and of the influence of system parameters such as the delay threshold. It is possible, in this way, to infer the effect of strategic decisions for performance assessment in transportation contracts. Such a parameter is also included in Carey's stochastic model (1999), but its evaluation is only possible under known distributions of primary delays. The delay threshold is also considered by Landex (2008), but this model does not allow detailed analysis of such parameter, as it ignores recovery from running time supplements and, so, the recovery region cannot be defined on the dimension of stations.

The novelty of this analytical approach is the ability to account explicitly for the running time supplement and the headway buffers at the same time, and the inclusion of further timetable parameters, the delay threshold under which delays are ignored. Previous approaches only considered delay propagation to the following trains, leaving out the spatial dimension (Carey, 1999; Carey and Kwieciński, 1994; Huisman and Boucherie, 2001; Landex, 2008; Zhu and Schnieder, 2000). The functional relation found in Landex's unbounded recovery model is, therefore, quadratic because delay propagation is only considered on the dimension of trains. Similar to this paper, Landex (2008) aggregates timetable slack to connect heterogeneous traffic to an ideal representative homogeneous condition. The weighted average proposed in this paper reveals more accurate than Landex's arithmetic mean of the headway buffers because the influence of timetable slack to recovery is now related to the distance from the primary delay location. Other approaches (Mattsson, 2007) incorporated only the recovery of individual trains along their trajectory and did not propagate delays to following trains.

Timetable slack has been integrated into models by indirect timetable metrics related to the ability to recover, such as density, flow, design speed (Hasegawa et al., 1981), and capacity consumption (Gibson et al., 2002). These metrics are often difficult to obtain in the planning phase and do not provide a clear picture of the possible individual train recovery.

Similarly to the model presented in our paper, Pyrgiotis (2012) proposed an airport network model to propagate delays over the air traffic. Railway traffic is, though,

much more constrained than air traffic, and in most railway lines the train sequences do not change between stations. This type of constraints cannot be implemented in Pyrgiotis's model, where arrivals and departures are assumed independent, and approaching queue capacity is assumed infinite. Timing points on railway lines often lay at simple halts with only one track per direction, where the train sequence remains strictly the same between arrivals and departures. Pyrgiotis applied the queuing theory to model the interferences between aircraft, whereas the model proposed here considers explicitly the headway buffer between trains. The two models are alike in the explicit formulation of the delay reduction given by the running time supplement along the train path or aircraft roster. However, the model presented in this paper provides also individual train delays as a linear function of primary delays, which allows to include infrastructure constraints, discontinuities in traffic volume or timetable parameters that cannot be described by a transposition of Pyrgiotis' model to railway networks. The availability of individual train delays opens further possibilities for new performance measures, based on this delay propagation model. Similarly to other stochastic models, Pyrgiotis' does not provide a functional relation between primary and cumulative delays and does not include the timetable parameter of delay threshold δ. Queuing theory models might still find application in the study of larger primary delays in converging networks, where the train sequence is not necessarily kept at the junctions, and dispatching criteria are required.

Furthermore, the polynomial formulation proposed in this paper provides insight into the relationship between primary and secondary delays. This is not possible using other delay models proposed in the past because the primary delays are accounted for implicitly in the variability of process times in real operation (Huisman and Boucherie, 2001; Pyrgiotis, 2012). Huisman and Boucherie focus on the secondary delays induced by the speed differences in the timetable, more than the relationship between primary and secondary delays. Huisman and Boucherie model railway operation in absence of a timetable, deriving the distributions of actual running times from the distributions of free running times and the actual buffers. The choice of queuing models is convenient in stages of planning when the timetable is not available, as also supported by Meester and Muns (2007). Other past models for delay propagation do not provide a functional relationship between primary and secondary delays (Gibson et al., 2002; Goverde, 2010, 2007; Meester and Muns, 2007). In particular, Meester and Muns use a recursive model of delay propagation, where every iteration depends directly and only on the previous iteration, so a functional relationship between primary delays and aggregate delay cannot be derived.

In the model presented in this paper, every individual train delay is expressed as a function of the primary delay, and relative location from primary delay, so perturbed conditions are always fully determined. The use of more complex methods, such as Colored Petri Net, has been also explored, but the same complexity represents the main downside of the model, which can represent effectively only a few stations (Zhu and Schnieder, 2000).

3.1.7 Conclusion

This paper contributes to the literature an analytic closed form function that returns individual secondary delays and cumulative railway line delay as a function of single or multiple initial primary delays. The function can predict the delays with high confidence, thus offering a fast analytic alternative to resource-consuming simulation models as demonstrated in the empirical analyses in section 5. The polynomial function may thus be used for an initial screening of possible timetables, leaving simulation to later parameter fine-tuning of timetable slack and delay threshold. Timetable optimization models might also benefit from this formulation integrating it in the objective function.

Most of the previous approaches for railway line delay propagation only considered alternatively running time supplements or headway buffers, resulting in lower degree functional relations. Others considered timetable slack in implicit form, as the difference between scheduled and maximum speed and traffic flow and did not support multiple primary delays and railway networks. Available Event Graph-based approaches do return the total delay in response of a set of primary delays, but the typical recursive approach makes it impossible to establish a functional relationship, which is available in the model presented in this paper.

Operation design tools such as the delay threshold, running time supplement, and headway buffer, can thus be designed accurately investigating the expected cumulative delay with an analytical approach. Differential calculus of the polynomial form shows that a limited amount of timetable slack is effective, whereas larger slack does not contribute to performance improvement and results in extending scheduled running times and delay recovery times.

The empirical tests in section 3.1.5 showed that the polynomial function model is robust to violations of the basic assumptions, and the form holds valid with heterogeneous running time supplements and headway buffers, provided that traffic is homogeneous.

The model returns also individual train delays as a linear function of a primary delay, which accommodates further measures of railway performance that might be

introduced in future research. Further development of the model should deepen the application to heterogeneous timetables and investigate the effect of the assumption of unvaried train sequences in merging networks. The closed-form model could be inverted to calculate average timetable supplement and buffer time from simulation results. This means that given a desired punctuality and stability of service, the necessary timetable supplement may be estimated from this function.

The universal formulation introduced here is non-specific to stations and trains, which allows analyzing the effects of primary delays occurring at any location on the railway line, and at any time of operation. The model is flexible and the aggregate delay resulting from different perturbations can be calculated efficiently with simple changes in the summation limits, keeping the same summation term. The recursive application of this model on homogeneous subsections of the railway line is suitable to estimate delay propagation on railway networks or on heterogeneous lines. Further, the model supports multiple primary delays at different trains and locations of the line. Thanks to the closed formulation, it is possible to quickly evaluate the effect of different strategic choices on contract performances between operators and transport authorities.

ACKNOWLEDGMENT: This work was funded by a Dean Grant from the Technical University of Denmark (DTU) and by the Danish Innovation Fund through the IPTOP project (Integrated Public Transport Optimisation and Planning).

References

Barron, A., Melo, P., Cohen, J., Anderson, R., 2013. Passenger-Focused Management Approach to Measurement of Train Delay Impacts, in: Transportation Research Board, 92nd Annual Meeting. Transportation Research Board of the National Academies, pp. 46–53. doi:10.3141/2351-06

Caimi, G., Burkolter, D., Herrmann, T., Chudak, F., Laumanns, M., 2009. Design of a railway scheduling model for dense services, in: Networks and Spatial Economics. Springer US, pp. 25–46. doi:10.1007/s11067-008-9091-6

Carey, M., 1999. Ex ante heuristic measures of schedule reliability. Transp. Res. Part B Methodol. 33, 473–494. doi:10.1016/S0191-2615(99)00002-8

Carey, M., Kwieciński, A., 1994. Stochastic approximation to the effects of headways on knock-on delays of trains. Transp. Res. Part B 28, 251–267. doi:10.1016/0191-2615(94)90001-9

Cerreto, F., 2015. Micro-simulation based analysis of railway lines robustness, in: 6th International Conference on Railway Operations Modelling and Analysis. International Association of Railway Operations Research, Tokyo, pp. 164-1-164–13.

European Performance Regime project, 2013. Handbook for the European Performance Regime (EPR) Guidelines for actual and potential users. Vienna.

Gibson, S., Cooper, G., Ball, B., 2002. Developments in transport policy: The evolution of capacity charges on the UK rail network. J. Transp. Econ. Policy 36, 341–354.

Ginkel, A., Schobel, A., 2007. To Wait or Not to Wait? The Bicriteria Delay Management Problem in Public Transportation. Transp. Sci. 41, 527–538. doi:10.1287/trsc.1070.0212

Goverde, R., 2010. A delay propagation algorithm for large-scale railway traffic networks. Transp. Res. Part C Emerg. Technol. 18, 269–287. doi:10.1016/j.trc.2010.01.002

Goverde, R., 2007. Railway timetable stability analysis using max-plus system theory. Transp. Res. Part B Methodol. 41, 179–201. doi:10.1016/j.trb.2006.02.003

Goverde, R., Hansen, I.A., 2013. Performance indicators for railway timetables, in: 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings. IEEE, pp. 301–306. doi:10.1109/ICIRT.2013.6696312

Goverde, R., Meng, L., 2011. Advanced monitoring and management information of railway operations. J. Rail Transp. Plan. Manag. 1, 69–79. doi:10.1016/j.jrtpm.2012.05.001

Haith, J., Johnson, D., Nash, C., 2014. The case for space: the measurement of capacity utilisation, its relationship with reactionary delay and the calculation of the capacity charge for the British rail network. Transp. Plan. Technol. 37, 20–37. doi:10.1080/03081060.2013.844906

Harker, P.T., Hong, S., 1994. Pricing of track time in railroad operations: An internal market approach. Transp. Res. Part B 28, 197–212. doi:10.1016/0191-2615(94)90007-8

Hasegawa, Y., Konya, H., Shinohara, S., 1981. Macro-Model on Propagation-Disappearance Process of Train Delays. Railw. Tech. Res. Institute, Q. Reports 22, 78–82.

Hofman, M., Madsen, L., Groth, J.J., Clausen, J., Larsen, J., 2006. Robustness and Recovery in Train Scheduling - a simulation study from DSB S-tog a / s. 6th Work. Algorithmic Methods Model. Optim. Railw.

Huisman, T., 2002. Forecasting delays on railway sections. Adv. Transp. 13, 779–786.

Huisman, T., Boucherie, R.J., 2001. Running times on railway sections with heterogeneous train traffic. Transp. Res. Part B Methodol. 35, 271–292. doi:10.1016/S0191-2615(99)00051-X

Jensen, L.W., 2015. Robustness indicators and capacity models for railway networks. Technical University of Denmark.

Jensen, L.W., Landex, A., Nielsen, O.A., Kroon, L.G., Schmidt, M., 2017. Strategic assessment of capacity consumption in railway networks: Framework and model. Transp. Res. Part C Emerg. Technol. 74, 126–149. doi:10.1016/j.trc.2016.10.013

Landex, A., 2008. Methods to estimate railway capacity and passenger delays. Technical University of Denmark (DTU).

Mattsson, L.-G., 2007. Railway Capacity and Train Delay Relationships. Crit. Infrastruct. Adv. Spat. Sci. doi:10.1007/978-3-540-68056-7_7

Meester, L.E., Muns, S., 2007. Stochastic delay propagation in railway networks and phase-type distributions. Transp. Res. Part B Methodol. 41, 218–230. doi:10.1016/j.trb.2006.02.007

Nash, A., Huerlimann, D., 2004. Railroad simulation using OpenTrack. Comput. Railw. IX 45–54.

Parbo, J., Nielsen, O.A., Prato, C.G., 2016. Passenger Perspectives in Railway Timetabling: A Literature Review. Transp. Rev. 36, 500–526. doi:10.1080/01441647.2015.1113574

Parbo, J., Nielsen, O.A., Prato, C.G., 2014. User perspectives in public transport timetable optimisation. Transp. Res. Part C Emerg. Technol. 48, 269–284. doi:10.1016/j.trc.2014.09.005

Pyrgiotis, N., 2012. A Stochastic and Dynamic Model of Delay Propagation Within an Airport Network For Policy Analysis. Massachusetts Institute of Technology.

Rete Ferroviaria Italiana, 2015. Prospetto Informativo della Rete 2014.

Salido, M.A., Barber, F., Ingolotti, L., 2012. Robustness for a single railway line: Analytical and simulation methods. Expert Syst. Appl. 39, 13305–13327. doi:10.1016/j.eswa.2012.05.071

Salido, M.A., Barber, F., Ingolotti, L., 2008. Robustness in railway transportation scheduling, in: Proceedings of the World Congress on Intelligent Control and Automation (WCICA). IEEE, Chongqing, China, pp. 2833–2837. doi:10.1109/WCICA.2008.4594481

Schittenhelm, B.H., 2013. Quantitative Methods for Assessment of Railway Timetables. Technical University of Denmark.

Schittenhelm, B.H., 2011. Planning With Timetable Supplements in Railway Timetables, in: Annual Transport Conference at Aalborg University. trafikdage, Aalborg, DK.

TfL Investment Programme Management Office, 2008. Business Case Development Manual.

Toletti, A., 2016. Modelling customer inconvenience in train reschedul- ing. 16th Swiss Transp. Res. Conf. (STRC 2016).

Törnquist, J., 2007. Railway traffic disturbance management-An experimental analysis of disturbance complexity, management objectives and limitations in planning horizon. Transp. Res. Part A Policy Pract. 41, 249–266. doi:10.1016/j.tra.2006.05.003

UIC, 2004. Leaflet 406 - Capacity.

UIC, 2009. Leaflet 450-2 Assessment of the performance of the network related to rail traffic operation for the purpose of quality analyses - delay coding and delay cause attribution process.

Vromans, M.J.C.M., 2005. Reliability of Railway Systems. Netherlands TRAIL Research School.

Zhu, P., Schnieder, E., 2000. Modelling and Performance Evaluation of Railway Traffic Under Stochastic Disturbances. IFAC Proc. Vol. 33, 303–312. doi:10.1016/S1474-6670(17)38163-6

3.2 Paper III: Delay Estimation on a Railway-Line with Smart Use of Micro-Simulation

Cerreto, Fabrizio, Steven Harrod, and Otto Anker Nielsen. "Delay Estimation on a Railway-Line with Smart Use of Micro-Simulation." Edited by Gianluca Dell'Acqua and Fred Wegman. *Transport Infrastructure and Systems*, 2017, 867–74. https://doi.org/10.1201/9781315281896-112.

Abstract

This paper formulates a delay propagation model that estimates total railway line delay as a polynomial function of a single primary delay. The estimate is derived from a finite series of delays over a horizon that spans two dimensions: the length of the railway line and the number of trains in the service plan. The paper shows that the total delay estimate is a cubic relation for small primary delays. A probabilistic approach is presented to combine the total delay functions of primary delays given to different trains. The final estimate is the total delay on railway lines after a random incident has occurred. The model can be integrated into railway timetable analysis to reduce the number of necessary simulations and can be used when the computation speed is an issue, such as on-line rescheduling algorithms. The model is demonstrated with an analysis of a Danish suburban railway.

KEYWORDS: *Railway delay; Timetable quality; simulation*

3.2.1 Introduction

Operational stability and robustness are crucial for railway transport. Not only are the passengers or users of the service sensitive to these measures of quality, but railways are usually integrated systems or networks, and failures at one location of the system affect other locations and services, sometimes quite catastrophically. Railway network planners are faced with many decisions about what quality of service to provide and what resources to allocate to deliver this service. Much of the literature demonstrates that there are frequently multiple feasible alternatives to allocate resources, and each alternative has a unique performance profile with characteristic statistics, especially with regards to punctuality and robustness. The analysis of these alternatives frequently requires laborious and inconclusive modeling with simulation software.

This paper contributes to the literature with a closed form function estimate of the aggregate railway line delay propagation in response to a primary delay. Many railway and transit services are of the form of a single terminating railway line, and this function may supplement or replace the application of simulation for the exploration of alternatives. On many railway lines, passenger traffic is distributed over the line destinations, and aggregate delay is an appropriate measure of system performance and customer service.

This formulation is closed form under a set of assumptions and is later shown to be robust to variance. The formulation is derived from a finite series of deviations from the service plan (secondary delays) caused by a singular initial disruption (primary delay). The total delay generated by disruptions on a railway line depends on the interactions between the trains, and a different total delay function is derived for each scheduled train. The probabilistic approach presented in this paper allows estimating the contribution of the individual trains to the general function of the total delay on a selected railway line.

Using microsimulation, the model can be shown to be robust to deviations in assumptions, and the results may be used to establish bounds of the expected performance of simulation models, and thus reduce the use of simulation models in preliminary, exploratory studies. Railway microsimulation is known for its heavy computational requirement, and the models proposed in this paper introduce new estimation of the total delay on railway lines with a very limited used of microsimulation, restricted to the initial calibration phase.

3.2.1.1 Literature survey

Prior literature on operational stability and delay propagation may be classified as proposing parametric methods, providing analytical methods, or demonstrating applications of simulation.

Parametric measures are functional relations fitted to empirical or experimental data. In these measures, the cause and effect relationships may not be clearly understood, or it may be strongly limited to selected environments. Krueger (1999) presents many capacity estimation functions proposed for and validated on North American long distance railways. International Union of Railways (2004) defines procedures using timetable compression to estimate the capacity of European high-density railway lines. Gorman (2009) fits linear multiple regression functions to large data sets of train operations on a North American railway to estimate delay as a function of train planning decisions.

Analytical methods derive system performance measures from known or presumed cause and effect relations in the railway service plan. Among these, Hasegawa et al. (1981) applies a hydrodynamic analogy to model railway traffic. The study models the delay propagation as a shock wave in a compressible fluid and finds the total delay as a cubic function of the primary delays by means of propagative velocity. Harker and Hong (1990) estimate the delay on a mixed single and double track railway where the train path is not defined in advance and is subject to a stochastic dispatching decision en route. Higgins et al. (1995) formulate decision rules for operation on a single track railway and then calculate in closed form the expectation of system delay given a traffic pattern and defined probabilities of delay for trains, track segments, and terminals.

Railway delay models often lead to innovations in mathematics, such as Meester and Muns (2007) application of phase-type distributions. Meester and Muns derive the net delay distribution on connected railway network segments given the distribution of primary delays on each segment. The derivation asserts that recursive calculation of the solution may be attained with just three operations: sum, nonnegative excess beyond a bound, and maximum. The paper states that a phase-type distribution, a distribution of the absorption time of a continuous time Markov chain, can be contained in the three operations in closed form. However, the method depends on the assertion of independence of the primary delays. The method is demonstrated for a sample network of 24 directional line segments with seven transfer points.

Goverde (2010) presents an efficient delay propagation algorithm where timetables are modeled as timed event graphs (using max-plus algebra) and initial delays

are known. The algorithm is very fast and in a few seconds can calculate the delay propagation over a large network consisting of many interdependent services, such as the Dutch national railway timetable. However, the model offers no functional relationship, and results must be calculated for each scenario separately. Kroon et al. (2007) proposes a stochastic linear program for the optimal allocation of supplement time along the route of a train path and finds that in a variety of realistic scenarios the supplement time should not be allocated uniformly along the train path. Finally, and most closely related to this paper, Landex (2008) proposes a delay propagation model computing the transfer of delay between trains through the scheduled buffer times. This model is used to study the relationship between capacity consumption and the development of the disruptions but does not take into account the recovery of train delays according to the timetable allowance. Cerreto (2016) extends Landex's delay propagation model to include the timetable allowance. The total delay on a railway line is described as a composite polynomial function of the primary delay generated at a station, which is cubic for small primary delays. The model allows to calculate the total delay with a limited use of micro-simulation but returns a different total delay function depending on the first train delayed.

Simulation is widely used, experimentally and in practice. Relevant publications include Lindfeldt (2015), which extensively applies RailSys commercial railway simulation software to a variety of capacity and delay propagation topics. In particular, Lindfeldt simulates 336 timetable scenarios and then applies linear regression to determine the significance of many common heterogeneity measures in predicting aggregate secondary delay. Lindfeldt finds that the *mean pass coefficient*, a measure of the frequency of meets and overtakes, is the most significant indicator. Mattson (2007) uses microsimulation to study the interferences between trains under different capacity utilization values: Mattson finds this to be the most precise way to analyze secondary delays, but it is also demanding for very detailed input and the process is very time-consuming. Lastly, Cerreto (2015) applies OpenTrack commercial railway simulation software to the analysis of a 21 km. line (nine stations) in the Netherlands with four configurations ranging from double track to quadruple track. Cerreto investigates methods to reduce the computation necessary for simulation-based analyses and limits the number of simulation runs required with a heuristic process called the *skimming method*. Instead of simulating all combinations of trains and delays, a composite profile of train delay is estimated from an initial simulation analysis, and this composite delay function is used to calculate the aggregate system delay. The results demonstrate that capacity utilization is

not strongly correlated with aggregate secondary delay, which contradicts the findings of some other literature.

3.2.2 Incident, primary delay probability, and total delay

Delays are positive deviations between the realized times and scheduled times of activities. In the literature, different classifications of delays are available. Most of the classifications distinguish between delays that are due directly to the variability of process times and delays that are originated by the subsequent conflicts in the actual operation (Goverde and Hansen, 2013). The *primary delays* are unexpected extensions of the planned times of the individual processes scheduled. For instance, equipment failures and large passenger flows generate primary delays. The *secondary delays*, on the other hand, are delays generated by operation conflicts, which are due to primary delays themselves. When a train is delayed, it needs to use infrastructure elements at different times than planned. A conflict arises when two or more trains request to use the same element at the same time: they will be queued by dispatching decisions since only one train at a time can use one element or track section. The delay that generates from the queuing is called secondary delay.

The *cumulative delay*, or *total delay*, on a railway line is the sum of all the total positive deviations registered for all the trains at all the time measurement points.

The delay generation process begins with a disruption or incident. A primary delay generates when the failure intersects a scheduled event in the timetable, and secondary delays evolve from the interaction between different scheduled events in the timetable.

Figure 3.2-1 - Delay generation process: Primary delays happen when incidents cross scheduled events. Secondary delays generate from delayed scheduled event crossing other events in the timetable.

The model presented in this paper translates the probability density distributions of incidents on a railway line into the probability densities of primary delays and of secondary and cumulative delays.

The section below describes the probability of generation of primary delays to a selected train, given the characteristics of the incident and the timetable.

3.2.2.1 Probability of primary delay to one train

We consider those incidents that prevent trains from moving. Such events can be, e.g. failures at signal boxes, extended boarding times at stations, failures at other ground or onboard systems.

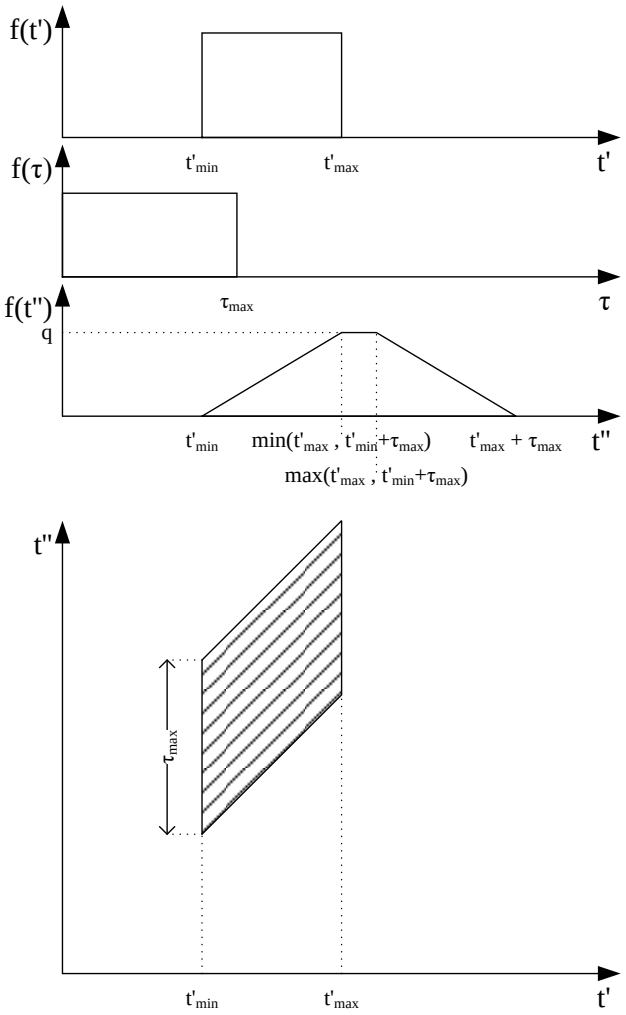

Figure 3.2-2 – Probability density functions of the starting time, the duration, and the ending time of an incident. The last graph shows the joint probability density domain of t' and t''.

In several cases, it is possible to describe an incident by the distributions of its starting time and duration. In the first instance, we adopt the relaxed assumption of uniform distributions in a given range. The formulation is, thus, simplified, but could be integrated with specific distributions fitted to the given incidents. The distribution of the starting time

could be assumed uniform in preliminary studies when detail information is not available. The distribution of the incident duration is still subject of studies in the railway field. Meng and Zhou (2011) propose the Normal distribution to model the disruption duration on single track lines, while the Exponential distribution is used by Schranil & Weidmann (2013) in Switzerland; finally Zilko et al (2016) propose an online model to predict the duration of a failure, based on the available knowledge at the beginning of the failure. The model uses the Copula Bayesian Networks to estimate the contribution of given influencing factors, based on historical data. In early studies the information available on the incidents may be insufficient to estimate these distributions, so we take the relaxed assumption of uniform distribution also for the incident duration for a simpler formulation. Our model translates the probability of incidents into the probability of primary delays by integration of the probability densities. The structure of the model would not be affected by choosing different distributions of the incident durations.

We define t' the starting time of an incident, t'' its ending time, and τ its duration, so that $t'' = t' + \tau$. Both t' and τ are assumed uniformly distributed, on independent ranges:

$$t' \epsilon\, \mathcal{U}(t'_{min}, t'_{max})$$

$$\tau \epsilon\, \mathcal{U}(0, \tau_{max})$$

Consequently, t'' follows a trapezoidal distribution in $[t'_{min}, t'_{max} + \tau_{max}]$ (Figure 3.2-2). The central segment of the distribution spans from $\min\{t'_{max}, \tau_{max}\}$ to $\max\{t'_{max}, \tau_{max}\}$, and its constant value is $q = \frac{2}{t'_{max} - t'_{min} + \tau_{max} + |\tau_{max} - t'_{max}|}$.

We define L_i the event "*Train* i *experiences a primary delay"*. The departure time of train *i* from the considered station is named θ_i, and the time separation between the train *i-1* and the train *i* is the headway $h_i = \theta_i - \theta_{i-1}$. The incident generates a primary delay to the train *i* if it starts between the departures of trains *i-1* and *i*, and it ends after the scheduled departure of the latter, θ_i.

$$L_i = \left(t' \in (\theta_{i-1}, \theta_i)\right) \cap (t'' > \theta_i)$$

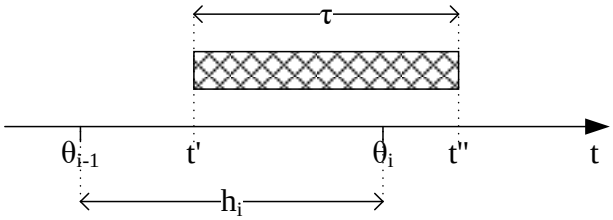

Figure 3.2-3 - A train receives a primary delay if the incident begins (t') in the previous headway (h_i) and it ends (t'') after the scheduled departure time (θ_i).

The intersection probability is expressed by means of the conditional probability:

$$P(L_i) = P(t'' > \theta_i \mid t' \in (\theta_{i-1}, \theta_i]) \cdot P(t' \in (\theta_{i-1}, \theta_i]) \quad (1)$$

t'' depends on *t'* through τ, and the conditional probability is derived hereunder.

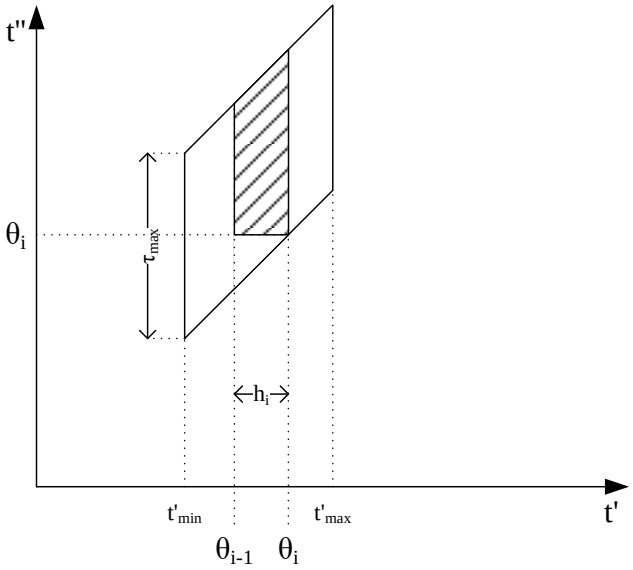

Figure 3.2-4 - Joint conditional probability $P(t'' > \theta_i \mid t' \in (\theta_{i-1}, \theta_i])$.

The conditional probability density of $(t'' \mid t' \in (\theta_{i-1}, \theta_i])$ has a trapezoidal shape in the range $[\theta_{i-1}\ ,\ \theta_i + \tau_{max}]$, with a central constant segment in the range $[\min\{\theta_i, \tau_{max} + \theta_{i-1}\}, \max\{\theta_i, \tau_{max} + \theta_{i-1}\}]$, and height $q = \frac{2}{h_i + \tau_{max} + |\tau_{max} - h_i|}$. The joint conditional probability corresponds to the striped area in Figure 3.2-4.

In the following formulation, equation (1) is split into two factors for a simpler explanation. We name the conditional delay probability $P(E1_i) = P(t'' > \theta_i \mid t' \in (\theta_{i-1}, \theta_i])$ and the event probability $P(E2_i) = P(t' \in (\theta_{i-1}, \theta_i])$ that corresponds to the start of the incident between trains *i-1* and *i*. The probability of $E1_i$ depends on the relation between τ_{max} and h_i and is described by the following:

$$P(E1_i) = \begin{cases} 1 - \dfrac{h_i}{2 \cdot \tau_{max}} \; for \; h_i < \tau_{max} \\ \dfrac{\tau_{max}}{2 \cdot h_i} \; for \; h_i > \tau_{max} \\ \dfrac{1}{2} for \; h_i = \tau_{max} \end{cases}$$

P(E2$_i$) is proportional to the headway of the train in the timetable cycle:

$$P(E2_i) = \frac{h_i}{\sum_i h_i} = \frac{h_i}{c}$$

where *c* is the timetable cycle and is given by the sum of all the headways. The probability of every train in the cyclic timetable to experience a primary delay is given by the following:

$$P(L_i) = \begin{cases} \dfrac{h_i \cdot (2\tau_{max} - h_i)}{2 \cdot c \cdot \tau_{max}} \; for \; h_i < \tau_{max} \\ \dfrac{\tau_{max}}{2 \cdot c} \; for \; h_i > \tau_{max} \\ \dfrac{h_i}{2 \cdot c} for \; h_i = \tau_{max} \end{cases} \quad (2)$$

Note that the probabilities of the individual trains to receive primary delays do not sum up to 1. We denote *P(0)* the probability that no train is delayed, that is

$$P(0) = 1 - \sum_i P(L_i) \quad (3)$$

3.2.2.2 Combined total delay functions

A total delay function describes the relation between primary delays given to a train and their cumulative effect on the railway line. Different train paths in a timetable are characterized by different stopping patterns, running time supplements and headway buffer times towards the following trains. Therefore, at every train path scheduled corresponds a characteristic total delay function of the primary delay.

We combine the characteristic total delay functions of different trains in a general total delay function that represents the effect of a primary delay to any of the trains in the timetable. The general function is a weighted average of the individual curves, where the weights are proportional to the individual probabilities of the trains to receive a primary delay.

The general total delay function is expressed by

$$d = \sum_i w_i \cdot d(i) \tag{4}$$

with weights

$$w_i = \frac{P(L_i)}{\sum_j P(L_j)} \tag{5}$$

Equation (4) allows the estimation of the general total delay given by an aleatory incident through the combination of total delay functions generated by selected trains. The estimation of individual total delay functions is relatively simple. In the following section, we describe a delay propagation model to calculate the total delay *d(i)* as a cubic function of the primary delay given to train *i*.

We reduce considerably the simulations necessary to estimate the general total delay combining the model described below and the probabilistic approach.

3.2.2.3 A finite series model of the total delay as a function of the primary delay

Previous literature demonstrates that the total delay on a railway line can be described as a cubic function of the primary delays given to a train. Cerreto (2016) models the total delay from the service timetable at all measurement points, as a function of timetable supplement, timetable buffer, and a single initial delay to one train. The model is summarized in this section.

The total delay model has a two-dimensional analysis domain, namely the length of the line and the number of trains included in the cumulative delay statistic. Trains on a single line with a single direction of movement are considered, which is a common operating plan in Europe and urban North America. The time horizon of the model then begins with the departure of the first train at the beginning of the line and ends with the arrival of the last train at the end of the line.

The total delay d represents the unweighted utility loss experienced by the railway service due to a disruption. It is the sum of all individual delays at measurement points in the timetable over the analysis horizon and is presented in (6).

$$d = \sum\nolimits_{j,s} (d_{j,s} \mid d_{j,s} \geq 0) \tag{6}$$

with $d_{j,s}$ being the delay of train j registered at station or timing point s (the difference between real and scheduled time).

The individual train delay $d_{j,s}$ is a combination of the hindrance from previous trains and the residual delay from the previous station. The delay is transferred to following

trains due to a lack of buffer time, while a train keeps a residual delay from the previous station due to a lack of running time supplement. Equation (7) expresses the delay propagation on the two dimensions of the model, under the relaxed assumption of equal running time supplement *a* for all the trains between any pair of stations and equal buffer time *b* between any pair of trains.

$$d_{j,s} = p - (s-1)a - (j-1)b \tag{7}$$

Subject to the non-negativity constraint: $d_{j,s} \geq 0 \; \forall \; j, s$.

p is the primary delay, which corresponds to the first train's delay at the first station $d_{1,1}=p$.

The total delay is derived summing up the individual train delays at all the stations. It results in (8).

$$d = \sum_{j,s|d_{j,s}>0} d_{j,s} = \sum_{s=1}^{\frac{p}{a}} \sum_{j=1}^{\frac{p-(s-1)a}{b}} p - (s-1)a - (j-1)b = \frac{(a^2+3ab)}{12ab}p + \frac{a+b}{ab}p^2 + \frac{1}{6ab}p^3 \tag{8}$$

The equation is valid for small values of primary delay that expire before the last train and before the last station.

Cerreto validates the model using microsimulation on a Danish suburban railway line with a heterogeneous timetable. The model is robust and holds valid when the assumptions of equal running time supplement and buffer times are removed. The total delay on the line can be regressed to a cubic polynomial function. The application to a heterogeneous timetable, though, returns a different cumulative delay function for each train that receives a primary delay.

We introduce the index *i* to identify the total delay function *d(i)* resulting from a primary delay given to train *i.*

The general total delay function is derived in section 3.2.2.2 combining the individual functions through the probability of each train to receive a primary delay.

3.2.3 Case study: The Nordbane in Copenhagen

We simulated the operation of a suburban railway line in Denmark to validate the combination of different polynomial functions to describe the total delay against the primary delay. The suburban railway network in Copenhagen is a very densely occupied network with 2 minutes headway in the busiest section. Six different lines operate on the

network, five running on the same central section. The suburban line is operated by uniform rolling stock in a cyclic timetable. The selected section of the suburban network is the line from Hellerup to Hillerød. Overtakes in this section are prevented. Though it is theoretically possible at selected stations, it hardly occurs in real operation, due to the very high frequency of the train service.

The micro-simulation software OpenTrack by OpenTrack Railway Technology Ltd. and the Swiss Federal Institute of Technology (ETH Zurich) was used for the simulation. This micro-simulation uses continuous computation of train motion equations and simulates the interaction between trains through discrete processing of signal box states (Nash and Huerlimann, 2004). Given user defined infrastructure, rolling stock, and timetable databases, it is possible to calibrate the train paths defining the running time supplements; moreover, different driving behaviors can be modeled for on-time trains and delayed ones. The strength of the micro-simulation models is the higher accuracy than the analytical models, and their flexibility to represent different contexts. Changes in the infrastructures and operating rules can easily be implemented and tested. The accuracy comes, though, at the cost of much longer computation time, as well as set-up time. Other micro-simulation software is available on the market, like RailSys by Rail Management Consultants GmbH (RMCon). Despite some differences in the approach, both the mentioned software suffer from long time needed to compute such detailed models (Landex, 2008).

Two different train paths run every ten minutes on the line between Hellerup and Hillerød with two different stopping patterns:

- Line A: runs throughout the entire line, skipping 5 stops in the first stretch
- Line E: only runs the first stretch, stopping at all the stations.

The line stationing and the schedules are summarized in Table 3.2-1.

The defined set of {1,…,10} minutes of primary delay was assigned separately to each train departing from Hellerup. The individual total delay functions were regressed from the corresponding total delay measured in the simulation, independently for line A and line E. The general total delay function of the line is calculated by the weighted average of the individual total delay functions of the trains.

Station		**Stationing**	**Schedule***	
Name	Code	km	A	E
Hellerup	Hl	7,8	05	07
Bernstorffsvej	Btf	9,3	|	09
Gentofte	Gj	10,9	|	11
Jægersborg	Jæt	12,6	|	14
Lyngby	Ly	13,9	11	16
Sorgenfri	Stf	15,9	|	19
Virum	VG	17,7	|	21
Holte	Hot	19,0	16	23
Birkerød	BG	23,8	21	
Allerød	LG	29,3	26	
Hillerød	HG	36,5	32	

*Table 3.2-1 – Line stationing and scheduled. *Departure minutes of the hour reported. Each train path repeats every 10 minutes. | = pass-through.*

For the model validation, we Monte Carlo sampled *n=200* failures at the departure signal from Hellerup, starting at a random time independent of the timetable. We regressed the measured the related total delay developed on the line to individual functions for every delayed line. The starting time of the disruption was extracted from a uniform distribution between 0 and 80 minutes, spanning over 8 consecutive timetable cycles. The duration of the failure was extracted from a uniform distribution between 0 and 10 minutes.

Table 3.2-2 compares the cases of primary delay experienced by each train line and the calculated probability. The weights for the general total delay function are calculated from the modeled probabilities.

Course	**Cases of Recorded primary delay**		**Model probability of primary delay**	**Weight**
(i)	#	%	$P(L_i)$	w_i
0	76	38.0 %	34.0 %	
A	91	45.5 %	48.0 %	0.73
E	33	16.5 %	18.0 %	0.27

Table 3.2-2 – Cases of primary delay registered in the simulation and probabilities modeled.

The total delay general function of the railway line was regressed from the whole set of simulations and compared to the combination of the individual delay functions.

Figure 3.2-5 compares the modeled general total delay on the line and the measured general total delay from the simulation.

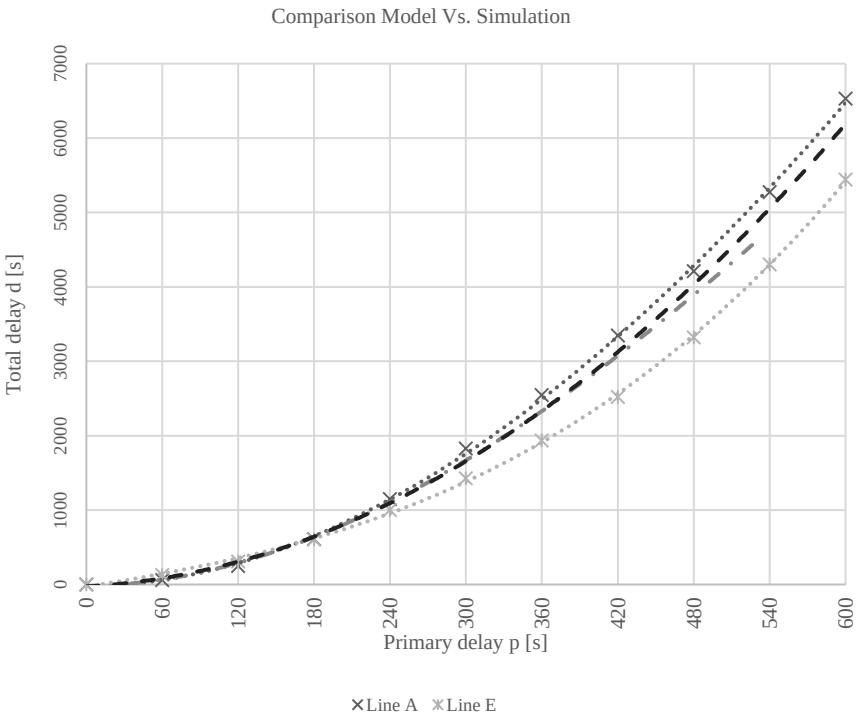

Figure 3.2-5 - Total delay on the line as a function of the primary delay given to Line A (dotted dark gray line). and Line E (dotted light grey line). Modeled (dashed black line) and measured (dot-dashed line) general total delay.

3.2.4 Results and discussion

The total delay on a railway line can be regressed to a cubic function of the primary delay. Every train that receives the primary delay returns a different function, due to a different interaction with the following trains, i.e. different buffer time.

The weighted average total delay function reflects the total delay function given by the joint simulation of failures independent of the timetable. In this case study, a series of 200 microsimulation of a random failure at a signal box was well approximated by a reduced series of 20 microsimulations of primary delays to individual trains.

The modeled total delay function and the measured total delay hold tight up to 500 s of primary delay. This is due to a higher number of trains from line E that received

smaller primary delay. As opposed, trains from line A tended to be the first delayed trains with higher values of primary delay. For this reason, the joint regressed total delay function is closer to line E for small primary delays and closer to line A at higher values of primary delay. The modeled general total delay function, instead, keeps the distance ratio between the two individual total delay functions throughout the entire primary delay range. A solution to this issue could be to cluster the distribution of the failure duration. In this way, different probabilities to be delayed can be calculated for different ranges of primary delays. The averaging weights would be calculated for individual clusters and the general total delay function would adapt to the probability to be delayed of individual trains.

3.2.5 Conclusions

This paper derives the total delay on a railway line as a closed function of the primary delay, under the assumption of equal buffer time between trains and equal running time supplement over the line.

We turn the estimation of the total delay given by an aleatory incident into the combination of the total delay functions of different trains. We determine each function's contribution to the general total delay with a probabilistic approach. The individual total delay function of each train is regressed from microsimulation. The result is a combined total delay function that does not depend on what train receives the primary delay. It is now possible to estimate the consequences of a given incident, simulating independent primary delays on the individual trains instead. This allows broader timetable analyses without increasing the number of simulations needed.

The model allows to calculate the total delay on a railway line with high accuracy from the microsimulation, reducing the amount of simulation runs needed. Using this model, we only needed one-tenth of the microsimulations used to estimate the total delay from the incident distributions. The number of microsimulations needed for the analysis may be further reduced, taking advantage of the good regressions of the individual total delay functions.

This is relevant for railway planners because it allows timetable accurate analyses with a limited computational power or on extended railway networks. At the same time, the accuracy of the model, together with the reduced computation needs, allows new applications in real-time rescheduling models, based on the total delay estimation.

The model accuracy could be further improved in the future clustering the distribution of the incident duration and introducing more complex distributions of the incident duration and starting time.

3.2.6 References

Cerreto, F., 2016. A Cubic Function Model for Railway Line Delay, in: Spessa, E., Biscotto, M., Smyrnakis, G. (Eds.), TRAVISIONS 2016 Contest Book. Warsaw, pp. 68–69.

Cerreto, F., 2015. Micro-simulation based analysis of railway lines robustness, in: 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015). International Association of Railway Operations Research, Tokyo, Japan, pp. 164-1-164–13.

Gorman, M.F., 2009. Statistical estimation of railroad congestion delay. Transp. Res. Part E Logist. Transp. Rev. 45, 446–456. doi:10.1016/j.tre.2008.08.004

Goverde, R.M.P., 2010. A delay propagation algorithm for large-scale railway traffic networks. Transp. Res. Part C Emerg. Technol. 18, 269–287. doi:10.1016/j.trc.2010.01.002

Goverde, R.M.P., Hansen, I.A., 2013. Performance indicators for railway timetables, in: 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings. IEEE, pp. 301–306. doi:10.1109/ICIRT.2013.6696312

Harker, P.T., Hong, S., 1990. Two Moments Estimation of the Delay on a Partially Double-Track Rail Line with Scheduled Traffic. J. Transp. Res. Forum 31, 38–49.

Hasegawa, Y., Konya, H., Shinohara, S., 1981. Macro-Model on Propagation-Disappearance Process of Train Delays. Railw. Tech. Res. Institute, Q. Reports 22, 78–82.

Higgins, A., Kozan, E., Ferreira, L., 1995. Modelling delay risks associated with train schedules. Transp. Plan. Technol. 19, 89–108. doi:10.1080/03081069508717561

Kroon, L.G., Dekker, R., Vromans, M., 2007. Cyclic railway timetabling: A stochastic optimization approach, in: Geraets, F., Kroon, L., Schoebel, A., Wagner, D., Zaroliagis, C.D. (Eds.), Lecture Notes in Computer Science. Springer, pp. 41–68. doi:10.1007/978-3-540-74247-0_2

Krueger, H., 1999. Parametric modeling in rail capacity planning, in: WSC'99. 1999 Winter Simulation Conference Proceedings. "Simulation - A Bridge to the Future" (Cat. No.99CH37038). IEEE, pp. 1194–1200. doi:10.1109/WSC.1999.816840

Landex, A., 2008. Methods to estimate railway capacity and passenger delays. Technical University of Denmark (DTU).

Lindfeldt, A., 2015. Railway capacity analysis - Methods for simulation and evaluation of timetables, delays and infrastructure. KTH Royal Institute of Technology.

Mattsson, L.-G., 2007. Railway Capacity and Train Delay Relationships. Crit. Infrastruct. Adv. Spat. Sci. doi:10.1007/978-3-540-68056-7_7

Meester, L.E., Muns, S., 2007. Stochastic delay propagation in railway networks and phase-type distributions. Transp. Res. Part B Methodol. 41, 218–230. doi:10.1016/j.trb.2006.02.007

Meng, L., Zhou, X., 2011. Robust single-track train dispatching model under a dynamic and stochastic environment: A scenario-based rolling horizon solution approach. Transp. Res. Part B Methodol. 45, 1080–1102. doi:10.1016/j.trb.2011.05.001

Nash, A., Huerlimann, D., 2004. Railroad simulation using OpenTrack. Comput. Railw.

Paper III: Delay Estimation on a Railway-Line with Smart Use of Micro-Simulation IX 45–54.

Schranil, S., Weidmann, U., 2013. Forecasting the Duration of Rail Operation Disturbances, in: TRB 92nd Annual Meeting Compendium of Papers. Transportation Research Board of the National Academies, Washington, D.C., pp. 1–20.

UIC, 2004. Leaflet 406 - Capacity.

Zilko, A.A., Kurowicka, D., Goverde, R.M.P., 2016. Modeling railway disruption lengths with Copula Bayesian Networks. Transp. Res. Part C Emerg. Technol. 68, 350–368. doi:10.1016/j.trc.2016.04.018

4 DATA ANALYSIS OF THE REALIZED OPERATION

4.1 Paper IV: Causal Analysis of Railway Running Delays

Cerreto, Fabrizio, Otto Anker Nielsen, Steven Harrod, and Bo Friis Nielsen. "Causal Analysis of Railway Running Delays." In *World Congress on Railway Research (WCRR)*, 1–7. Milan, Italy: World Congress on Railway Research, 2016.

Abstract

Operating delays and network propagation are inherent characteristics of railway operations. These are traditionally reduced by provision of time supplements or "slack" in railway timetables and operating plans. Supplement allocation policies must trade off reliability in the service commitments against service transit times and railway asset productivity. Methods to investigate the quality of supplement time allocation are necessary to reduce the behavioral response and the waste of resources.

This is a preliminary study that investigates train delay data from the year 2014 supplied by Rail Net Denmark (the Danish infrastructure manager). The statistical analysis of the data identifies the minimum running times and the scheduled running time supplements and investigates the evolution of train delays along given train paths.

An improved allocation of time supplements would result in smaller overall aggregate timetable supplement, reduced transport travel times, and higher productive utilization of train rolling stock. The study results will lead eventually to both better allocation of time supplements in timetable structures, and identification of areas that should be a high priority for correction.

KEYWORDS; *Express trains; Punctuality; Railway; Statistics; Timetable Supplement*

4.1.1 Introduction

The railway industry commonly benchmarks itself through key performance indicators such as punctuality and reliability. These compact measurements express the quality of the service, meant as the ability to respect the schedule promised to the passengers. The running time supplement is one of the timetabling tools used to improve punctuality. This paper gives an overview of the running time supplement design and use in operation. It also presents a statistical approach to analyze historical data of train timekeeping in Denmark, in order to investigate the quality of the timetable supplement allocation. The purpose is to present different strategies for the design of timetable supplements and to assess their impact on punctuality.

With the objective of evaluating the effectiveness of the slack currently scheduled in train paths, this paper proposes statistical methods to quantify the running time supplement and compare it with the delay evolution through the paths. It is possible to identify areas where the running time supplement is not used and therefore wasted, and sections of the train paths where delays are not recovered, suggesting a lack of running time supplement.

4.1.1.1 Punctuality, primary delays, and secondary delays

Punctuality and delays are well known general concepts, but their definition and computation method vary among countries and railway companies. *Punctuality* refers to the number of trains that are not delayed, compared to the total number of trains operated (Olsson and Haugland, 2004). It can be attributed to individual stations or trains over a period of time, or it can measure railway networks entirely or partly. Differences are found in the selection of the punctuality measurement points and of the trains to be included in the measure. Accordingly, also the punctuality targets are different in every country and can be train category-specific (Schittenhelm, 2011). For example, punctuality is measured along the entire train path in Denmark, while only selected stations are counted in the Netherlands, Switzerland, and Germany. Other countries measure punctuality only at the final destinations, like Italy and Norway. It is common to differentiate the punctuality target between passenger and freight trains. In several countries, passenger trains are further divided into long distance and regional/suburban trains.

Delays are positive deviations between the realized times and scheduled times of activities. In the literature, different classifications of delays are available. Most of the classifications distinguish between delays that are due directly to the variability of process times and delays that are originated by the subsequent conflicts in the actual operation

(Goverde and Hansen, 2013). The *primary delays* are unexpected extensions of the planned times of the individual processes scheduled. For instance, equipment failures and large passenger flows generate primary delays. The *secondary delays*, on the other hand, are delays generated by operation conflicts, which are due to primary delays themselves. When a train is delayed, it needs to use infrastructure elements at different times than planned. A conflict arises when two or more trains request to use the same element at the same time: they will be queued by dispatching decisions since only one train at a time can use one element. The delay that generates from the queuing is called secondary delay (Cerreto, 2016). Dispatching decisions are crucial for the management of the delay propagation: Olsson and Haugland (2004) found that the dispatchers tend to use defined priority rules on single tracked lines or in cases of large delays. Personal judgment prevails, on the other hand, on double-tracked lines or with small delays.

4.1.1.2 Timetable supplement

Scheduled times are usually longer than the minimum time required by processes. The difference between the scheduled times and the expected minimum realization times is referred to with different names by authors: *slack time*, *timetable allowance*, or *time supplement*. The timetable supplement is a tool that planners include in the timetables to compensate for natural variations of process times. It reduces the probability of generating primary delays, and it is expected to increase punctuality. On the other side, the supplement increases the traveling time and operating costs, resulting in a reduction of attractiveness and efficiency. To be effective and efficient, the timetable supplement should be properly dimensioned and distributed. Some strategies to allocate the supplement times are described below.

4.1.1.3 Allocation strategies for the timetable supplement

The allocation of the time supplement is a tradeoff between attractive travel times and timekeeping. General guidelines, built on empirical studies, are provided by the International Union of the Railways (UIC, 2000). The guidelines provide a fixed supplement to include in the train paths, proportional to the path length and increasing with the maximum speed, but they give no indication about the optimal distribution of the supplement along the paths. In addition, the recommendations are not mandatory and only suggest a minimum amount of supplement. Every railway planner has its own strategy to allocate the slack in the timetable and most western European countries use larger values then recommended. For example, the Danish railway Infrastructure Manager, RailNet Denmark, uses a flat distribution of the supplement on the regional and long-distance trains,

which, in some cases, doubles the UIC-recommended values. Condensation and compensation, instead, is the Swiss strategy for timetable supplement allocation. The network is divided into zones according to the capacity utilization. The capacity bottlenecks areas are named *condensation zones*, where the supplement time is minimized to reduce the capacity utilization. In contrast, large supplement times are scheduled in the areas that are not capacity bottlenecks, called *compensation zones*, to recover possible delays accumulated in the previous condensation zones (Schittenhelm, 2011).

The national strategies for the supplement time allocation typically reflect the way the punctuality is measured: for instance, Denmark measures punctuality at all the stations and spreads the supplement along the train paths, except in the Copenhagen suburban railway network. Switzerland measures punctuality at larger stations and concentrates the supplement before those stations. Norway measures punctuality at the final destination and schedules large amounts of supplements in the last segments of the paths.

4.1.1.4 Effects of the time supplement

A properly designed time supplement should lead to a better regularity of the scheduled process, improving the railway punctuality. The relation between supplement time increase and punctuality improvement, though, is not straightforward. Carey (1998) formulated a behavioral response model to describe an observed phenomenon that reduced the benefit of supplement times. The main finding was that if more time is allowed to a process, the process self-adapts to the new schedule and takes a longer time on average. Train drivers tend to act slower in the departure procedures and to drive slower, passengers tend to take longer to board and alight, dispatchers tend to use the extra elasticity given by supplement times for train prioritization and delay management. In this sense, the supplement time could be thought as the capacity buffer between consecutive productive processes, which absorbs the inherent variabilities in the production. The risk is to hide systematic failures in the process, which should be tackled individually to increase the reliability. The famous case of the Sunset Limited train in the USA is reported by Larson (1998): the train schedule included such a large slack time that it had been hiding wrong dispatching strategies for years, and was consistently attaining poor punctuality. Adding even larger supplement times did not improve the train punctuality, while the increased travel time reduced the attractiveness for passengers.

Carey's theoretical formulation (1998) finds a balanced supplement time allocation optimizing the total cost, which consists of the cost of the scheduled trips and

the cost of the expected delays. The cost of the scheduled trips is proportional to the trip length, so it is minimized with short running times and, therefore, minimum running time supplements. The cost of the expected delays decreases non-linearly enlarging the supplement times. A reduction in the expected delay is mirrored by a relevant reduction in costs for fuel, equipment utilization, and overtime wages, as also mentioned by Johnston (2008).

4.1.2 Case study

New methods to design and allocate the running time supplements are subject of several studies with different methodologies. Our current research focuses on the statistical study of historical data to assess whether the timetable supplement in existing timetables fits the actual need and if it is properly used.

In the following subsection, we present methods to investigate the actual use of the time supplement in train paths and compare it to the scheduled timetables through the statistical analysis of historical data from the daily operation.

4.1.2.1 Minimum running times

As described in the previous sections, the scheduled process times consist of the minimum process time and a slack time, or time supplement, to absorb inherent variations of the process time. Therefore, the planners need to compute the minimum running time between two stations. Different tools support this operation, each of them with a different approximation. Acceleration and deceleration models can provide approximated running time estimation, especially on simple plain lines. Micro-simulation of train motion allows a more accurate computation. It and can easily be combined with detailed infrastructure models to take into account slopes and the train's tractive effort and braking power (Cerreto, 2015). Real tests on the lines can be performed running trains on free tracks, but this type of tests is expensive and hard to realize. Each estimation method has its own uncertainties that should be evaluated.

We used historical data from RailNet Denmark from 2014, third quarter, to investigate the realized running times in the past. The actual minimum running times were identified on the railway line Copenhagen – Roskilde, the most congested line in Denmark. The investigation covers only the express trains ("*Lyntog*") that stopped at the bigger stations. The scheme below outlines the 30 km long line and the stopping locations.

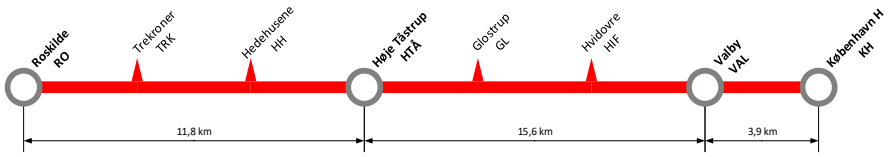

Figure 4.1-1 - Railway line Copenhagen - Roskilde. Express trains only stop at the major stations.

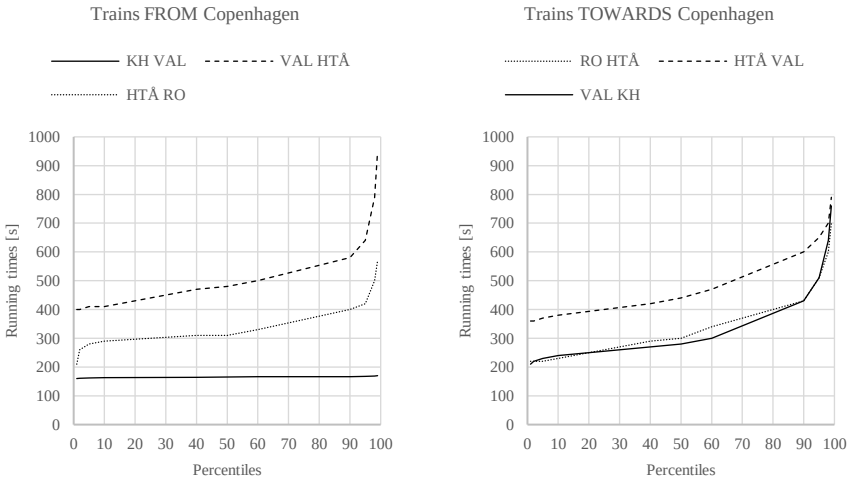

Figure 4.1-2 – Actual running time percentiles of the express trains on the railway line Copenhagen – Roskilde, divided by segment and direction.

The charts above represent percentiles of the actual running time distributions, divided by segments between stops, and by direction. The scheduled supplement time was filtered by referencing the minimum running time at the second percentile of the distributions. The second percentiles filtered well also running times that were too short, possibly due to the accuracy of the recordings or to random errors.

Differences in the distributions of the two directions are worth further investigation. The spread of running times by segment is considerably wider for trains from Copenhagen. The segment closest to Copenhagen changes significantly in stability between the two directions, being almost constant for trains from Copenhagen. The future investigation could highlight the existence of a behavioral response to supplement time allocated at the departure, as Copenhagen is often the origin of long-distance trains.

The 2014 schedule varied considerably over the day, even for trains from the same category and scheduled with the same stopping patterns and rolling stock. The changes made it not possible to identify a unique running time supplement for each leg. Further research will investigate the variability of allocated supplement over the day. The supplement time will be estimated for individual trains through longitudinal statistics over the whole year.

4.1.2.2 Delay, delay variation and supplement times

Alongside the minimum running times, we compared the train delays at different stations to evaluate the delay development.

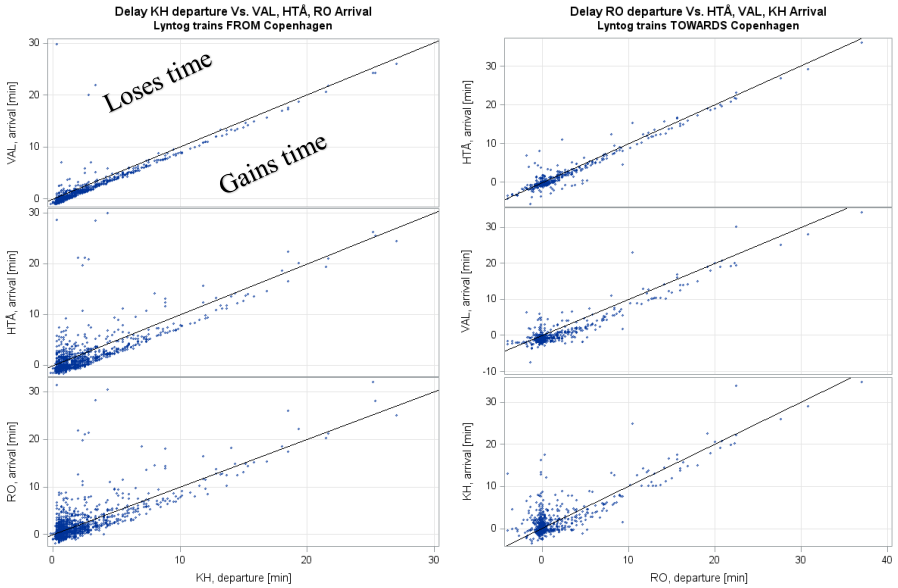

Figure 4.1-3 – Departure delays at the beginning of the line compared to arrival delays at the following stations. Reference lines drawn at equal delays.

For both the directions, the delay departing from the first station was compared to the arrival delay at the last three stations. A reference line is set to $x=y$ in the plots, where x is the departure delay at the first station and y is the arrival delay at the following stations. Points below this line represent trains catching up their delay, while points above the reference line mean that the trains increase their delay along the way. For both the directions, the charts follow the train path top-down. The majority of the points lay near the reference line, indicating natural variations in the delays that normally occur over 30 km of line.

Differences in the cloud density are visible between the two directions: while trains from Copenhagen are tight, trains bound for Copenhagen spread wider over the charts. The same phenomenon is visible in the comparison of delays from trains traveling in the same direction. The vertical distance between the individual points and the reference line represents the train's recovery or loss. The trains leaving Copenhagen show a clear recovery pattern at Valby, shaping a straight line parallel to the reference line. This shape fades out at Høje Tåstrup and Roskilde, with more dispersed points showing more variability. An explanation is found in the distance covered by trains, as mentioned by Olsson & Haugland (2004): the section of the line that we considered is the final segment of many long distance trains bound for Copenhagen and the initial one for trains from Copenhagen. For this reason, trains to Copenhagen are subject to higher variability in delays. The realized recovery on one section could be modeled, thus, as an aleatory variable. The charts show that the realized recovery on consecutive line sections does not sum up linearly, but as aleatory distributions. A model is worth deeper investigation and theoretical formulation.

Early departures from Copenhagen are forbidden, as visible in the scatter plot. On the other side, some express trains that do not stop at Roskilde are allowed to travel early at this station. The right-hand scatter plot shows that the earliness of several trains at Roskilde translates into late arrivals in Copenhagen. An extension of Olsson and Haugland findings on dispatching decision (2004) could suggest the prevalence of dispatchers' personal judgment also for early trains and should be further investigated.

The association of the higher running times registered for trains from Copenhagen, and the variation in delay recovery between Høje Tåstrup and Roskilde, suggests the existence of scheduled supplement times that are not used to recover delays. This excess should be quantified to optimize the resources utilization in future timetables. On the other hand, early trains to Copenhagen, traveling out of their designated slot, could relate to an excess of supplement time in the section before Roskilde. An optimal distribution of the supplement time should prevent excessive earliness, reducing dispatching issues at the bigger nodes, and resulting in better punctuality.

4.1.3 Conclusions

This paper reports the preliminary results of a research on train delays under development at the Technical University of Denmark, within the research project IPTOP (Integrated Public Transport Optimization and Planning).

Today's access to large-scale data makes it possible today to apply multivariate statistics to the recordings of railway operation, based on automated train detection systems.

Previous studies identified several influencing factors in punctuality. Nevertheless, new methods to identify excessive and insufficient timetable supplements are necessary. This paper shows that the actual supplement time can be detected in a train path by means of historical data. Further, the possibility to spot delay and recovery patterns is presented, and the impact of dispatching strategies will be developed in future research.

Recurring delay patterns may be found dependent on the infrastructure layout, the rolling stock performance and reliability, the time of the day and of the year, and the stationing on lines and at stations. Delay causes tracking is regulated under the UIC leaflet 450-2 (2009), which sets a standard codification, thus the structure of this analysis is applicable in many nations. The availability of detailed information on delay causes will also offer the possibility to deepen the previous studies on punctuality influencing factors. Delay causes recording is now required for international trains by the International Union of Railways and it is also being adopted for national trains among the railway infrastructure managers, giving access to data unavailable before. Primary and secondary delays should be explicitly recorded, in this way, making it possible to develop algorithms to link primary and secondary delays, and to further clarify how trains may auto-correlate their delays.

References

Carey, M., 1998. Optimizing scheduled times, allowing for behavioural response. Transp. Res. Part B Methodol. 32, 329–342. doi:10.1016/S0191-2615(97)00039-8

Cerreto, F., 2016. A Cubic Function Model for Railway Line Delay, in: Spessa, E., Biscotto, M., Smyrnakis, G. (Eds.), TRAVISIONS 2016 Contest Book. Warsaw, pp. 68–69.

Cerreto, F., 2015. Micro-simulation based analysis of railway lines robustness, in: 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015). International Association of Railway Operations Research, Tokyo, Japan, pp. 164-1-164–13.

Goverde, R.M.P., Hansen, I.A., 2013. Performance indicators for railway timetables, in: 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings. IEEE, pp. 301–306. doi:10.1109/ICIRT.2013.6696312

Johnston, B., 2008. Delays get attention in Congress. Trains 30.

Larson, J.L., 1998. The battle of the Sunset Limited. Trains 64–69.

Olsson, N.O.E., Haugland, H., 2004. Influencing factors on train punctuality—results from some Norwegian studies. Transp. Policy 11, 387–397. doi:10.1016/j.tranpol.2004.07.001

Schittenhelm, B.H., 2011. Planning With Timetable Supplements in Railway Timetables, in: Annual Transport Conference at Aalborg University. trafikdage, Aalborg, DK.

UIC, 2000. UIC leaflet 451-1 Timetable recovery margins to guarantee timekeeping - Recovery margins, 4th ed.

UIC, 2009. Leaflet 450-2 Assessment of the performance of the network related to rail traffic operation for the purpose of quality analyses - delay coding and delay cause attribution process.

4.2 Paper V: Application of Data Clustering to Railway Delay Pattern Recognition

Cerreto, Fabrizio, Bo Friis Nielsen, Otto Anker Nielsen, and Steven S. Harrod. "Application of Data Clustering to Railway Delay Pattern Recognition." Published in the *Journal of Advanced Transportation*, 2018, 1–18. https://doi.org/10.1155/2018/6164534. Presented at *RailCPH 2017* (Banekonferencen), Copenhagen, Denmark, May 15, 2017.

Abstract

K-means clustering is employed to identify recurrent delay patterns on a high traffic railway line north of Copenhagen, Denmark. The clusters identify behavioral patterns in the very large ("big data") data sets generated automatically and continuously by the railway signal system. The results reveal where corrective actions are necessary, showing where recurrent delay patterns take place. Delay profiles and delay-change profiles are generated from timestamps to compare different train runs and to partition the set of observations into groups of similar elements. K-means clustering can identify and discriminate different patterns affecting the same stations, which is otherwise difficult in previous approaches based on visual inspection. Classical methods of univariate analysis do not reveal these patterns. The demonstrated methodology is scalable and can be applied to any system of transport.

KEYWORDS: *Railway Delay; Big Data; K-means clustering; Historical data mining*

4.2.1 Introduction

Operations analysis is the collection and review of performance data, such as punctuality and process cycle time. It is a key step in the continuous improvement of transport services, and several methods exist to collect and analyze data from operations. The increasing availability of automated data sources is offering new ways to analyze operations, providing deeper insight and more reliable information. Railway management is very accepting of these new possibilities, and considerable effort is made by operators and institutions to use operations analysis in feedback loops for improving the timetabling process (D'agostino, 2016; Peterson, 2012; Richter, 2008; Schittenhelm and Richter, 2009). A better understanding of the development of delays in railways, and in transportation in general, provides the opportunity to improve the processes and identify the factors affecting reliability. For example, causes of delays might be identified in misallocation of supplements and buffers in timetables, structural conflicts that require mitigation actions, suboptimal design of station processes, and inefficient procedures for preparing a train for departure. This paper demonstrates a data-mining technique based on k-means clustering to identify recurrent delay patterns in transportation, identify the main reason for cluster membership, and provide managerial insight to improve timetables and processes.

Prior studies propose several methods that are currently in use for operation analysis, deploying sources of automatic data collection. These approaches can be divided into traditional statistical methods and big data techniques, which differ in both the use of data and in the output provided. Traditional methods tend to aggregate and summarize information, so these can provide a general picture or detailed information on specific stations or trains. These are typically proposed in the form of multiple univariate distribution analysis, where the occurrence of different delay patterns at the same station is not visible. Big data techniques can be used to investigate recurring patterns or internal structures in operations. These approaches are expanding, thanks to the growing availability of large amounts of data, and several techniques have been deployed to identify recurrences of delays and describe or predict delays. Advanced techniques such as neural networks, succession rules, Bayesian networks, and various methods of regression, have been developed mainly to predict delays real-time in railways, as described in §4.2.2. However, train delays are necessarily correlated over the progression of a complete journey, and these data relations both along the journey of a train and among adjacent train paths have not received as much attention in the literature.

This paper presents a big-data technique to identify recurring delay patterns in railway operations. Big data refer to information assets characterized by high volume, velocity, and variety, which value is extrapolated by analytical methods (De Mauro et al., 2016). In this application, the absolute delay and delay change are tracked for individual train paths along a railway line, resulting in absolute delay and delay change profiles. In the papers based on univariate statistics, systematic delays in these profiles are identified through visual inspection. The manual search for similarities suffers from subjective interpretation from the operator and is easily biased by common artifacts of the representation. The technique presented in this paper applies k-means clustering to find recurrent patterns in train delay progression, so that management may identify processes for improvement or correction. In this way, it is possible to support continuous quality improvement.

In the next section, §4.2.2, a literature survey of contemporary data analysis methods is offered. §4.2.3 presents the k-means cluster method and the structure of the data to be studied. §4.2.4 presents results from the study of a high-density Danish railway line. The effectiveness of k-means clustering for this application is discussed in §4.2.5, with particular regards of its novelty compared to existing literature, while the conclusions of this paper are presented in §4.2.6.

4.2.2 Literature survey

Operations analysis is fundamental in the continuous improvement process to manage and modify railway operations. Data collected from real operations, or from simulation models, has been used in the feedback loop to design and improve railway timetables for decades. Typically, even if timetables may change over time, some of the fundamental infrastructure and service behaviors will not be modified. Timetables are often the result of only minor modifications to the previous editions and need to consider problems discovered in earlier timetables. For example, after a structural change in the Danish railway timetable in 1998, after the opening of the Great Belt fixed link, the service structure remained largely unchanged until 2016 (Hansen, 2015).

Data collection systems have proliferated in railway networks since 2000, and very large amounts of data are available today. Widespread systems to collect data increased both the volume and the variety of data, which are often collected by different systems at the same time. The methods to elaborate and interpret information from past operations evolved together with the amount and quality of data, starting from descriptive and inferential statistic and moving towards big-data techniques. For example, delay

probability density functions can be extrapolated from historical data and integrated into analytical models to estimate service reliability before operation (Carey, 1999). Goverde et al. (2001) performs extensive statistical analysis and distribution fitting of data from the Dutch railway network. Goverde et al. fits different distributions for arrival and departure delays and finds that no general distribution fits all groups of recorded arrival delays.

Primary delay distributions derived from operational data are also often employed as input in simulation models to evaluate the propagation of delays. Sipilä (2010) explores the effect of modified running time supplements in railway schedules through microsimulation of a Swedish railway line. The author identifies different strategies for running time supplement allocation by verifying the significance of the change in punctuality recorded in 1600 simulations of selected scenarios. Olov Lindfeldt (2010) describes a method to aggregate delay data from real records and isolate distributions of primary delays. These distributions are then used to formulate microsimulation models. The data consists of manual records from dispatchers, who assign a delay cause code to every record greater than 4 minutes of delay on the Swedish railways. In absence of other sources of data, the reliability of manual record cannot be validated, although the whole simulation model and its results rely on the derived distributions. Studies from other countries show that manual input can be indeed unreliable (Goverde and Meng, 2011; Sørensen et al., 2017). The same method to extract primary delay distributions is later used by Anders Lindfeldt and Sipilä (2014) in a simulation model to assess the effect of allowing freight trains to travel outside of their assigned path. The authors demonstrate that the realized travel times of freight trains could be shortened considerably without affecting the performance of other trains. The reduction of unnecessary waits for traffic management, and the permission to depart before schedule reduces the average travel time on one side but increases its variability on the other.

Historical data also provides insight into the factors that influence service reliability. Olsson and Haugland (2004) apply regression analysis on the Norwegian railway network and identify the most relevant factors for punctuality, such as absolute passenger flow and passenger occupation ratio. Gorman (2009) uses regression analysis on data from American single-tracked freight railways to identify the factors that contribute the most in prolongation of railway running times. Gorman predicts congestion delay based on meets and passes scheduled as a consequence of speed heterogeneity. Again in simulation, Shih et al. (2014) applies an approach similar to Gorman's to determine the best capacity expansion strategy in terms of reduction of average

prolongation of running time for freight trains. Shih et al. identifies functional relationships, through regression of simulation results, between average delay per train-mile and several factors, such as the relative length of the double-tracked section of a railway line. Anders Lindfeldt (2010) applies multilinear regression with a special focus on F-statistics to investigate factors generating delays on the Swedish railway network. Lindfeldt measures delay changes over selected routes and analyzes their distributions. In particular, the response variables are the share of trains with a delay increase, the median change in delay, and its standard deviation on the route. Statistically significant explanatory variables are found in the traffic volume for both passenger and freight trains. Among passenger trains, the most significant variables are average speed and traffic heterogeneity, and for freight trains, it is the number of stations on the route with at least three tracks.

Time stamps and recorded deviations from schedule can be integrated with information from other sources. For example, incident reports may be compiled in case of larger disruptions. Such reports include information about the typology of the incident, the train affected by the primary delay, other trains involved, the secondary delays generated, and the recovery plans taken by the dispatchers. Schittenhelm and Richter (2009) describes the reporting system in the Danish railways (the same system in service at the time of this study) and introduces a quantile-based approach to depict the development of train delays en-route. The plots confirm the general understanding of delays from experienced operators and can be used to quantify the magnitude of expected disruption. The quantile-based approach, though, describes operations as a whole, and it is not able to distinguish systematic delays occurring at individual stations, but with different origins, so analysis of individual train services is necessary to identify peculiar delay patterns. Richter (2010) introduces new metrics to identify improvement actions, based on data from automatic detection systems. Richter sorts the trains according to recorded delay and identifies the worst in a percentile approach, associated with recorded delay causes. A similar approach is adopted with regards to change in deviation, or delay jump, recorded on line sections so that most critical geographical areas are identified. Lastly, Richter proposes a tabular representation of the median delay of individual trains recorded at the station, sorted by scheduled time and geographical location. In this way, the analyst can identify which specific trains typically suffer from primary delays, also characterized by geographical location, and which are the trains typically affected.

Similarly, Peterson (2012) studies the on-time performance along the path of specific train services, using the rolling average delay of the last three timing points. Such

on-time performance is plotted for all the repetitions of a specific train service over a time period, and compared to the average, standard deviation, and 75th percentile. Peterson identifies empty areas in the pool of plotted delay profiles and interprets these as recurrent delay patterns given by discrete dispatching choices along the train path. Peterson also interprets recurrent increases or decreases of vehicle delay as segments of insufficient or excess running time supplement, respectively. Reliability of service is described by the standard deviation of recorded delays. Peterson used the mentioned measures in a feedback loop to redistribute the running time supplement in train paths according to the recorded performance.

Andersson et al. (2011) assesses the effectiveness of running time supplement in railway schedules from empirical data collected on a Swedish railway line. The study plots the recorded delays over the train itinerary overlapped with the scheduled running time supplement and compares pairwise the stacked plots from different railway services, stopping patterns or directions. The identification of misallocation of running time supplement is based on a visual search for recurrent delay patterns, and a few different dispatching tactics are identified in clusters of similar delay profiles. Andersson et al. highlights the existence of a threshold value of delay that triggers prioritization of other trains that are traveling on schedule. The observations are clustered in groups and show recurrent delay patterns, and the analysis is supported by a detailed analysis of possible conflicts among individual train itineraries. Noticeably, the authors demonstrate that the measures of punctuality currently in use on the Swedish network hide the effects of running time supplement misallocation and delays developed en-route. Even though the punctuality at the final destination is a measure of railway performance very common among railway operators, it does not express how trains increase or recover from delays along their journey. Schittenhelm (2011) provides a sample of similar measuring approaches in the European railway industry. In a later study, Andersson et al. (2013b) underlines the relevance of critical points for network robustness by plotting delay profiles and showing that the profiles cluster around critical points according to different dispatching strategies. Advanced clustering techniques may support the identification of different strategies to compute the effects on robustness.

Lastly, van Oort et al. (2015) evaluates data collected automatically on public transport services with a combination of statistical methods and visual representation. The study represents delay data similarly to Peterson (2012), Andersson (2013b, 2011), and Schittenhelm (2009), plotting the recorded delay over individual repetitions of the same

service path, and adds the plot of relevant delay percentiles over the stations. The shape of the percentile-based delay profiles highlights recurrent patterns in the deviation from schedule. The representative delay profiles appear different depending on the percentile they represent. Patterns found included the presence of typical early arrival at stations in bus services, followed by waiting time until the scheduled departure time, or recurrent delay drops or increases at specific stations. The delay plots are combined with the measured headway from the previous vehicle. While the delay plots would suggest allocating more running time supplements at systematic increases of delay, structural delays that cannot be compensated by timetable slack are highlighted in the plots of headways, were service unreliability corresponds to scattered recorded headways. A percentile approach was also presented by van Oort et al. to characterize and sort the stations according to performance, similarly to previous literature.

The statistical analyses presented above are suitable for the general description of the system performance but lack specific insight on recurrent delay patterns that occur in operation, and on the relationships between delays at different locations. The literature presented in this survey focuses on the univariate analysis of selected measures, such as delays at single stations. Traditional metrics common in the railway industry, such as punctuality, have also been found unrepresentative of the actual service reliability. The methods that include the multidimensional aspect of the problem mostly deal with delay profiles, the sequences of delays recorded on individual train itineraries. The quality of these analyses often relies on visual inspection of plotted data, and the observer-operated search for matching delay profiles. This search lacks a standardized methodology and is influenced by the plotting layout and the subjective interpretation, which is based on personal experience.

Big data techniques have arisen recently and seek to make use of the very large amount of information that is provided by automatic data collection systems, overcoming the mentioned issues of traditional methods. The term big data is rather broad and includes different techniques that serve a specific purpose. The common characteristics of these techniques are Volume, Velocity, and Variety, meaning large amounts of data, generated at high speed, possibly by different sources with different or no structure (De Mauro et al., 2016). As opposed to standard statistical analyses, where hypotheses are formulated and tested, big data techniques search for internal structures directly in the data. Data generated by automatic sources typically fit into the big-data criteria. In railways, several data mining techniques were developed in the last years, following different approaches and searching

for different types of information. The interest in these techniques is rising, together with the increasing availability of structured data. Industrial applications of these techniques are spreading, and new approaches are being studied also among public institutions (D'agostino, 2016).

Event mining is a technique based on time sorted logs, where relations between different events are found based on their coincidences. Hansen et al. (2010) combines an event mining tool and standard statistics to predict the actual running times of trains to the next station, given all the recorded current delays. Dependencies between pairs of events are found or "mined" in timed event graphs created from the time stamps of individual trains, which correspond to events of occupation and release of blocking sections. The process times between events are inspected by standard statistics, resulting in conditional probabilities of process times, given the recorded delays of all relevant trains in the system. Such a model, though, relies considerably on very detailed knowledge about the infrastructure and requires data which is not commonly available from railway infrastructure managers.

Goverde and Meng (2011) uses the same information source and similar technique to identify and analyze route conflicts and identify delay chains. Infrastructure data and operation data are integrated so that it is possible to identify a train that is occupying a blocking section linked to a signal at danger for another train. Delay trees are built and traced backwards to identify the primary causes, so individual delays can be classified automatically into primary and secondary, and the correct attribution of delay causes can be verified. Interestingly, the authors verify that more than half of the delay-cause records were assigned wrongly by the dispatchers, stating that, in the Netherlands, this type of manual input is not reliable and objective enough to be deployed in data analysis.

Kecman and Goverde (2012) extends the model to include non-logged line sections, where it is not possible to distinguish delays due to signaling impositions and delays due to primary causes. Delay chains are also traced in less detailed data by Sørensen et al. (2017). Based on the time sequences at stations experiencing disturbed operations, the authors identify the trains generating the conflicts and the trains suffering from the conflicts. The analysis is used to identify primary delays, describe single days of operation, identify frequent trains originating, or subject to, delay chains, and identify point stations where most of the primary or secondary delays are generated. In a comparison with manually recorded delay causes, the study finds relevant inconsistencies with the primary

delays traced in the delay chains, in accordance with Goverde and Meng (2011). The method described, though, is only valid for single track lines and does not identify multiple primary delays.

Cule et al. (2011) introduces association rules to identify delays recurring often together and sets up an episode mining framework to highlight frequent delay patterns from train timestamps at stations. However, association rules can highlight common recurrences, but cannot explain relations of causality between two events, so primary and secondary delays cannot be distinguished. Similarly, Wallander and Mäkitalo (2012) identifies delay chains according to the manual delay cause records from the dispatchers and based on timestamps at stations with a granularity of 1 minute. The succession rules used are very similar to association rules, but consider the time dependencies, so that events taking place earlier can be assumed to be the cause of events happening later under the same circumstances. Trains are characterized by the number and magnitude of conflicts they generate so that improvement actions can be concentrated. Association rules have also been adopted to evaluate the effectiveness of delay prevention actions on Japanese suburban networks by Yabuki et al. (2015). Yabuki et al. compares the association among occurrence of delays of different trains, change in delays, extension of running and dwelling times and realized headway in before/after scenario comparison. The downside of such models is that association rules can be set between binary variables, so the development of delays depicted does not include its magnitude. Further, the number of associations to be analyzed grows exponentially with the number of potential pairs of events, so the analyses must be limited to short time frames of operation.

Neural networks are a big-data method that learns from historical records and uses the relations identified among variables to predict an output, given unseen values of the input variables. This technique is particularly suited to delay prediction and has been deployed in multiple studies. Neural networks look for dependencies in the data, as opposed to simulation models, which are based on interaction rules between objects defined initially. Malavasi and Ricci (2001) uses neural networks to predict the total experienced delay on a railway line, given its geometrical and technological characteristics, and its scheduled utilization over time. In comparison to simulation, Malavasi and Ricci find neural networks more robust against extreme-valued input, which implicates more likely case-overfitting with simulation. Kecman et al. (2015) proposes a Bayesian network delay prediction model. In this case, the input includes the timetable and recorded delays at all stations. Each delay is assumed to depend only on direct connections in a timed event

graph, meaning the recorded delay for the same train at a previous station, and for the previous train at the same station. Conditional delay distributions are assumed Gaussian, and the parameters are derived through recursive Generalized Linear Models. Chapuis (2017) deploys the same assumed delay dependency in a neural network model, where input includes the delay of the previous train and at the previous station, and distance to the next station. Such a model can predict the delay of a train at the next station. Independent of the actual infrastructure, this model is generic and can be applied at any station of the railway network. The downside of neural networks, though, is the risk of data overfitting, reducing the prediction capability, although this risk is lower in neural networks than in simulation models.

In response, Marković et al. (2015) introduces Support Vector Regression (SVR) to establish a functional relationship between the characteristics of the railway system and train delays. Train category, scheduled time, infrastructure, and share of the journey completed are identified as most influencing factors to predict the train delay at one station. The authors show that SVR generalizes better than an artificial neural network, which seeks to minimize the error of prediction in the historical dataset. Interestingly, the authors assume that the performance of delay prediction can be improved by grouping delays by magnitude, as factors generating smaller delays differ from factors that generate larger disturbances.

Kecman and Goverde (2015) applies big data techniques to predict running and dwelling times from actual operation data, based on records from block sections occupations. The study uses random forests of tree-based models, to predict non-linear relations between input variables and process times, with sufficient robustness to outliers in the data, lowered risk of overfitting, and with focus on real-time application. Running time predictors are calculated for every block section, and dwelling time predictors are calculated for every station platform. Among the interesting findings, the running times are longer if the headway to the preceding train is short, meaning that the succeeding trains tend to slow down to smoothen the trip and reduce the risk of encountering a yellow signal. Moreover, the authors find no evidence to support the hypothesis that trains run faster when delayed. All the trains were found to run at approximately the maximum performance in any condition. The authors suggest that, in case of insufficient prediction accuracy, new variables might be included in the model, such as the platform shape for dwelling times.

Big data techniques focus mainly on the prediction of delays and running times, or in the identification of delay chains and realized delay propagation among trains. New applications of these techniques would support the analysis of the realized development of delays along the path of individual train delays. As shown by statistical analysis and visual search for patterns presented by Schittenhelm and Richter (2010; 2009), Peterson (2012), Andersson et al. (2013b, 2011) and van Oort et al. (2015), this type of data contains a great deal of information yet to be explored, which would provide insight on the effectiveness of running time supplements, and on the presence of structural issues that generate delay in transport operation. In this paper we present a clustering technique to identify recurrent delay patterns among train services, based on readily available data, and which leaves room for inference on the factors that generate specific delay patterns. The result shows that, within comparable train trajectories and stopping patterns, different train services accumulate delay at different stations, and that recovery shapes differently according to the route direction. Inferences on the cluster composition show the most frequent service characteristics in each cluster. Such information could guide the allocation of corrective measures to improve timetables. Table 4.2-1 and Table 4.2-2 summarize the literature just reviewed.

	Environment		Technique					
	Real operation	Simulation	Distribution Fitting	Test significance	Regression analysis	Percentile sorting	Visual inspection	Purpose
Goverde et al. (2001)	X		X					Distributions of Primary and Secondary delays
Sipilä (2010)		X		X				Comparison running time supplement strategies
O. Lindfeldt (2010)	X	X	X					Distributions of Primary delays from real operation for simulation
Olsson and Haugland (2004)	X				X			Factors affecting punctuality
Gorman (2009)	X				X			Factors that generate delays on single track lines
A. Lindfeldt (2010)	X				X			Factors that increase delays in line segments
A. Lindfeldt and Sipilä (2014)		X		X				Travel times with different operation models, with/without free freight operation
Shih et al. (2014)		X			X			Factors affecting average delay per train-mile
Schittenhelm and Richter (2009)	X					X	X	Visual inspection of quantile-based representation of deviations and change in deviation
Richter (2010)	X					X		Delay tabular representation and sorting train service performance
Peterson (2012)	X					X		Rolling average delay for specific train services
Andresson et al. (2011)	X						X	Assessment of effectiveness of running time supplements
Andresson et al. (2013)	X						X	Identification of critical points for robustness
van Oort et al. (2015)	X						X	Delay profiles, headway profiles

Table 4.2-1: Review of previous uses of univariate statistics in railway operation analysis

	Technique								Level of detail			
	Event mining	Association Rules	Succession Rules	Neural Networks	Bayesian Networks	Random forests	Support Vector Regression	Clustering	Track sections	Station	Input	Purpose
Hansen et al. (2010)	X								X		Current delays of all trains	Prediction of running time to next station
Goverde and Meng (2011) Kecman and Goverde (2012)	X								X		Timestamps	Delay chains, Actual primary delay causes
Sørensen et al. (2017)	X									X	Timestamps	Delay chains on single track lines, actual primary delay causes
Cule et al. (2011)		X								X	Timestamps	Delay patterns
Wallander and Mäkitalo (2012)			X							X	Timestamps, delay causes from dispatchers	Delay chains
Yabuki et al. (2015)		X								X	Timestamps	Comparison of real scenarios
Malavasi and Ricci (2001)				X					X		Physical infrastructure and utilization ratio	Prediction of total realized delay on a network
Kecman et al. (2015)a					X					X	Current train delay, last delay at station	Delay prediction at next stations
Chapuis (2017)				X						X	Current train delay, last delay at station, distance	Delay prediction at next stations
Marković et al. (2015)							X			X	Infrastructure and train journey characteristics	Delay prediction at next stations
Kecman et al. (2015)b						X			X		Current traffic condition, actual train position, delays of the day	Running time and dwelling time prediction
Cerreto et al. (2018) (This paper)								*X*		*X*	*Timestamps*	*Recurrent delay patterns across stations*

Table 4.2-2: Review of previous uses of big data techniques in railway operation analysis

4.2.3 Identification of recurrent delay patterns using big data techniques

In this paper, a delay profile of a train run is defined as the set of recorded deviations throughout its path or a part of it, on a specific date. Note that deviation is reported as the time difference between a scheduled and a realized event, such as arrival, departure, or a nonstop timing point. Even though the delay is often used to refer to positive deviations, a delay profile can include null and negative values. A delay profile is a

powerful representation of operation and the comparison of several delay profiles along the same service path allows the identification of recurrent delay patterns and such a representation method has already been presented in the literature (Andersson et al., 2013b, 2011; Peterson, 2012; Richter, 2010; Schittenhelm and Richter, 2009; van Oort et al., 2015). Delay change, also called delay jump, is the difference in deviation between two consecutive stations, and represents the delay recovery or increase. Schittenhelm and Richter (2009) use this measure to assess delay increases or time gains between stations, and Goverde and Meng (2011) use it to identify delay chains in railway operation. We define a delay change profile of a train as the set of recorded delay changes along its path or a part of it.

A dataset of delay profiles consists of all the delay profiles recorded in a defined period, stacked together. Fields, or variables, of the dataset are the events at every station, whereas observations are individual train runs from a selected service. Such a dataset can refer to a specific train service or to several services following the same stopping pattern so that the fields can be aggregated. The first case is intended for infrequent services, typically long-distance trains, where every single service may have its own characteristics in terms of planned demand, scheduled rolling stock, or the time of crossing congested nodes. Suburban and regional railway services are often scheduled in constant stopping patterns at high frequency, and could, thus, be analyzed together, expecting characteristics of operation to be more homogenous across services. A dataset of delay change profiles is defined analogously to delay profile datasets, where the fields contain the change in deviation in place of the absolute deviation.

Previous research presented on delay and delay-change profiles interpret recurrent patterns by the visual search for similarities (Andersson et al., 2013b, 2011; Peterson, 2012; Schittenhelm and Richter, 2009). The systematic analysis of these two types of datasets through clustering algorithms allows the identification of patterns that are not necessarily visible, or that could be wrongly associated by subjective interpretation.

Clustering techniques partition a dataset into a collection of groups of similar observations. In this study, clustering is used to partition the datasets of delay profiles and identify train services that are candidates for identification of common causality. Inference on common factors appearing in observations clustered together facilitates the assessment of delay patterns in association to specific characteristics of a transport service, such as time of the day (peak/off-peak), day of the week, or equipment used. The clustering process is realized through measures of similarity between elements in the same cluster

and dissimilarity between elements from different clusters. Several methods and metrics are available to accomplish the task, suitable for different uses. Hierarchical algorithms proceed by splitting or merging observations recursively and are preferred when a nested structure is assumed in the clusters. In contrast, partitional algorithms do not impose a hierarchical structure and find all the clusters at the same time. K-means clustering is a partitional algorithm and was chosen due to its simplicity and frequent appearance in the literature (Jain, 2010).

K-means clustering is an iterative clustering process based on the identification of the mean element in each cluster. Every cluster is represented by its centroid, calculated as the average of the elements of the cluster, and every observation is assigned to the cluster corresponding to the closest centroid. Given a number k of initial centroids, the algorithm executes the following steps:

1. assign every element to the cluster with the closest centroid;
2. calculate the new centroids of all the clusters as the mean of the elements;
3. repeat until convergence, which is met when no element changes cluster between consecutive iterations.

This simple method requires three user-specified parameters, which might be hard to determine beforehand. The distance metric, the number of clusters *k*, and the cluster initialization. Euclidean distance is often used to determine the difference between observations, but other metrics are available, such as the L_1 distance (Kashima et al., 2008). The number of clusters *k* is the most difficult parameter to estimate, as there is no perfect mathematical criterion. The parameter *k* is typically determined according to available knowledge about the data or interpreting and evaluating the meaning of several independent partitions realized for different values of *k*. The initial centroids might influence the resulting clusters, so the initialization is often chosen among several independent partitions that result from sampling *k* initial centroids among the observations. The influence of initialization, however, generally diminishes with the dimensionality of the dataset (Jain, 2010).

A substantial contribution to the simplicity of the method is given by the required structure of the data. Contrary to observer-operated search, clustering methods rely on the numerical relations between variable values recorded across single observations. It is, thus, unnecessary for the clustering algorithm to preprocess the data and sort the recorded delays for every train/observation. In the method proposed in this paper, k-means clustering is applied to observations of a multidimensional variable, whose size corresponds to the

number of timing points of a fixed stopping pattern, where the fields contain the delays, or delay changes, respectively, recorded at the individual timing points. Every observation of this multidimensional variable is a vector and represents a single train run.

4.2.4 Case study: The Kystbane, Copenhagen

The *Kystbane* (Coastline) is a double-tracked railway in the Copenhagen region. It is one of the busiest railway lines in the network of Banedanmark, the Danish infrastructure manager, and it is operated to regional standards, with some international services. It is operated nearly entirely by DSB, the largest Danish railway undertaking, which runs three different service types. The timetable is cyclic, and the services operate different stopping patterns during the day, as illustrated in Figure 4.2-1.

- The Øresund trains ("ØK") run all day every 20 minutes on a limited section of the coastline, between Copenhagen and Nivå. These trains operate between Denmark and Sweden across the Øresund bridge, and stop at every station in Danish territory;
- The Regional trains ("ØP") run all day every 20 minutes as well, but they only operate in Denmark and run the whole coastline. These trains skip selected stops between Copenhagen and Nivå;
- Additional trains are operated in the morning and afternoon peak hours. The Rush hour trains ("ØD") operate every 20 minutes between Copenhagen and Helsingør, skipping other selected stops.

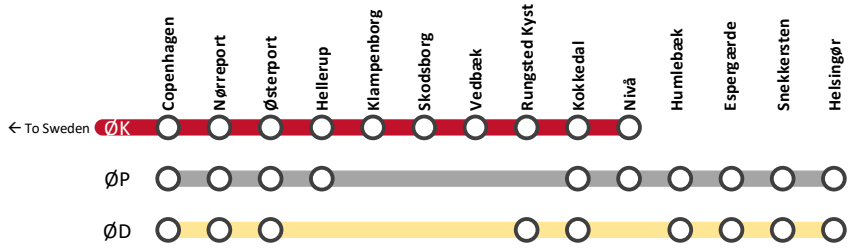

Figure 4.2-1: DSB services and stopping patterns on the Kystbane.

Fewer trains with modified stopping patterns run at night, so only weekday operation between 4:30 and 20:00 is considered in this study. The sections between Copenhagen and Østerport, and between Snekkersten and Helsingør are shared with other services and operators.

In the resulting charts, stations are identified by a code specified by the infrastructure manager. Station codes and names are reported in Table 4.2-3.

Station code	Station name	Distance from KH [km]
KH	København H (Copenhagen Central Station)	0,0
KN	Nørreport	1,5
KK	Østerport	3,1
HL	Hellerup	7,8
KL	Klampenborg	13,3
SÅ	Skodsborg	18,8
VB	Vedbæk	22,1
RU	Rungsted Kyst	26,1
OK	Kokkedal	29,1
NI	Nivå	32,5
HUM	Humlebæk	36,3
GÆ	Espergærde	40,0
SQ	Snekkersten	42,7
HG	Helsingør	46,2

Table 4.2-3: Station codes and names on the Kystbane

Banedanmark provided a set of timestamps that state the scheduled and realized times of the trains at every timing point from April to December 2014. The records include information about the operation and about the timing points, such as station name, train ID, train category, scheduled time and recorded deviation. Banedanmark relies on automatic train detection systems, based on the signaling system components. Typically, the track circuit boundaries do not correspond exactly to the platforms, and an offset is generated between the time recorded by the automatic system and the actual time a train arrives at the platform or departs. This is a rather common problem, and it is also reported in the Netherlands (Kecman and Goverde, 2015) and Norway (Sørensen et al., 2017). For the Danish network, a correction factor was calculated by Banedanmark using statistical analyses of GPS positions of train trajectories in collaboration with the main rail operator, DSB. The method and results are described by Richter et al. (2012; 2013). Nørreport station is the only station underground on the line, so GPS correction is not available, which is visible as a saw-tooth pattern common to all train services in the delay profiles presented below, with a slightly underestimated delay for arrival records at Nørreport and overestimated for departure records from the same station. Similarly, delay change records are shifted to negative values for arrivals at Nørreport, and at Østerport, whereas higher positive values are recorded for delay changes at departures from Nørreport. The bias is

systematic and has the same exact effect on all the trains, therefore its influence on clustering can be neglected.

The train time stamps were rearranged by an automatic algorithm to create datasets as described in §4.2.3, by means of the commercial software SAS 9.4 TS Level 1M4, by SAS Institute Inc., Cary, NC, USA. Observations corresponded to a realized train on a given date, and the fields contained the recorded delay at every station. Data from every station was divided in arrival, departure, and pass-through times, where trains did not stop. Each record is the delay profile or the delay-change of a train on a date and represents one observation of the given train. Every variable identifies the station code and the type of timestamp, which can be entrance to the station, I ("Indkørsel"), exit from the station, U ("Udkørsel"), or pass through station, G ("Gennemkørsel), which is used where trains do not stop.

The analysis is intended to report delay patterns. Consequently, punctual trains are discarded from the dataset. In Denmark, punctuality measurements are based on a delay threshold of 5 minutes for regional and long-distance trains, such as the Kystbane. However, for internal management purposes, the infrastructure manager Banedanmark creates a delay report every time a train reaches at least 3 minutes of delay, containing information on the delay cause and on possible other trains hindered. Consequently, only trains with at least one recorded delay greater than or equal to 3 minutes are considered relevant in the present case study. Delay distributions are known to include large shares of trains with short delays, with decreasing frequency for larger delays (Carey, 1999; Goverde et al., 2001). Largely unbalanced clusters are a known issue in clustering algorithms and are an object of study to reduce the interference of large clusters (Wu, 2012). In this case, punctual trains can, therefore, be considered as a compact cluster derived by prior knowledge, and they can be filtered out from the cluster analysis. The operation of filtering can be considered noise reduction and improves the quality of clustering, as the k-means procedure tends to generate spherical clusters of the same radius (Hastie et al., 2009). According to Marković et al. (2015), large delays are influenced by different factors other than smaller delays, which further supports the filtering choice. However, in different contexts, the filtering threshold might be set equal to a different value, or not be applied at all.

Given the characteristic high frequency of train services on this line, clustering was operated by stopping patterns rather than by train numbers, so trains were grouped together by direction and service category. Grouping trains with similar characteristics and

same stopping patterns increases data availability in the comparison and does not disqualify the result. In fact, such grouping was already proposed by Schittenhelm and Richter (2009).

As explained in §4.2.3, k-means clustering requires choosing the number of clusters *k* in advance. To set the number of clusters, the k-means algorithm was repeated with different values of *k*, and the best result was selected using criteria from Jain (2010). The number of clusters *k* should be large enough to represent different patterns. At the same time, as *k* increases, the same delay patterns tend to split into more clusters, and *k* should remain small enough to prevent the generation of duplicate clusters. In detail, for every combination of train category, direction, and clustering variable (delay or delay change), *k* was set as the highest integer that did not generate duplicate clusters. That is, the univariate distributions of delays, or delay changes, in every cluster should be different from all the other clusters for at least one station. Since *k* is selected independently for all the mentioned cases, the same set of trains might best be represented by a different number of clusters when the algorithm operates on the delay variables or on the delay change variables. The L1 distance was used as a clustering metric between observations, as suggested by Kashima et al. (2008).

K-means clustering was performed on the described dataset by the commercial software MATLAB R2017a, by The MathWorks, Inc.. In the following figures, selected results of the application of the method are reported, clustering on either delay profiles, or on the delay change profiles.

4.2.4.1 Clustering results

Figure 4.2-2 illustrates the effectiveness of delay profiles clustering on ØK southbound trains, on the delay variables. Note, after a stop at Copenhagen central station, these trains proceed to Sweden. The charts show that similar delay profiles are grouped together with low variance around the average centroid of each cluster, highlighting recurrent patterns. The resulting clusters can be interpreted as follows:

1. Cluster 1: Trains that are punctual on the first section of the line, but suffer delays approaching the most congested area of Copenhagen, mainly from Klampenborg and from Østerport;
2. Cluster 2: Trains that are punctual throughout the complete journey, which receive delays leaving from Copenhagen;

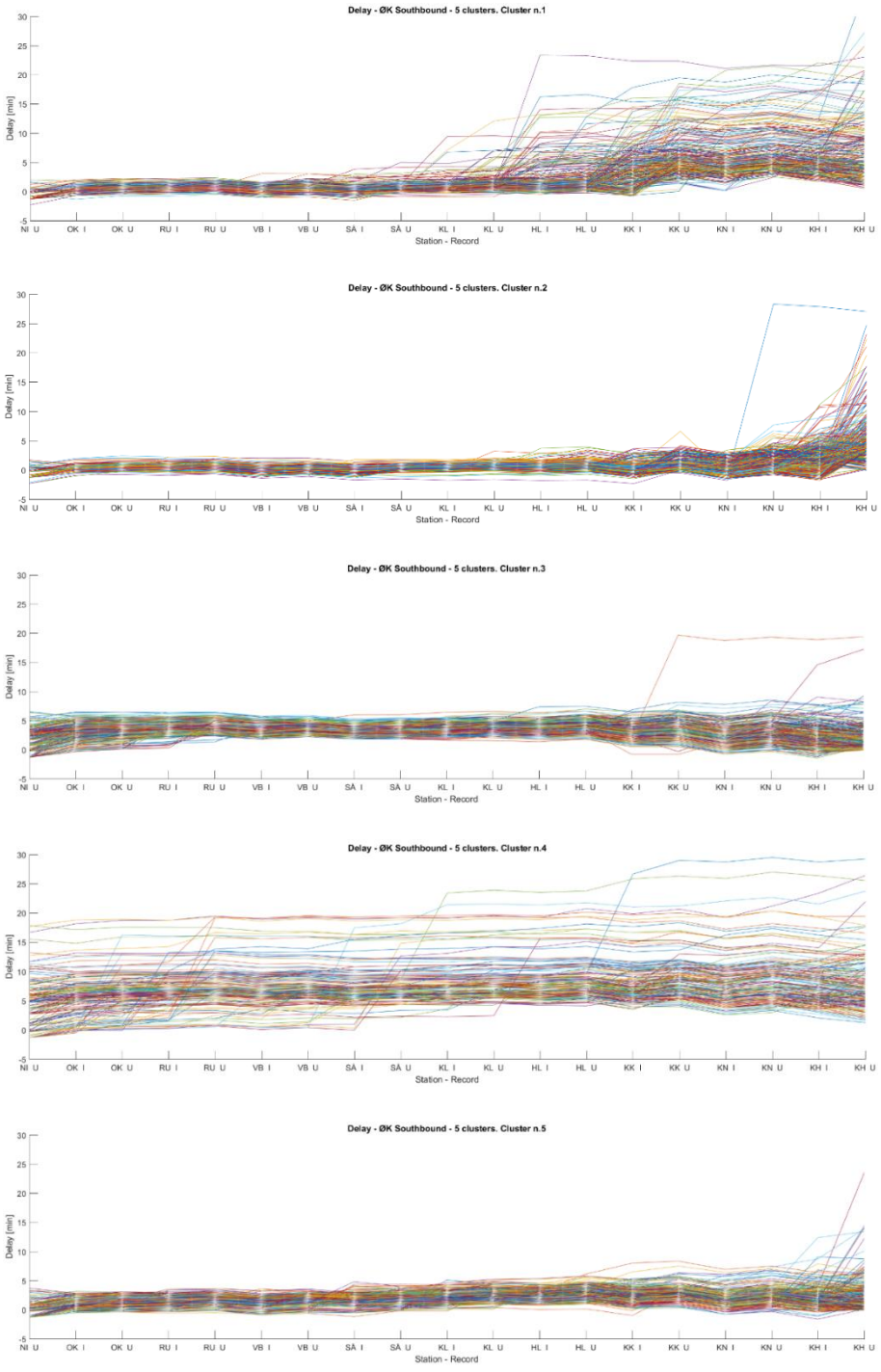

Figure 4.2-2: Resulting clusters in southbound ØK trains, Nivå – Copenhagen.

3. Cluster 3: Trains that are nearly punctual, or anyway within 5 minutes of delay through the complete journey, and across Copenhagen central station; from Hellerup, a marginal delay recovery is visible for these trains;
4. Cluster 4: The most delayed trains, being delayed throughout the whole itinerary, or it largest part;
5. Cluster 5: Punctual trains with slightly, but steadily, increasing delay across stations.

Some clusters present outliers, such as clusters 2 and 3. Even though some delay profiles may appear considerably different from other profiles in the same clusters, these observations were assigned to the cluster with the closest centroid. This means that, in selected cases, the delay profiles are the representation of rather unique events, which may be neglected after more detailed analysis in the composition of the individual clusters.

Individual clusters are characterized through the mean values of the aforementioned measures. The following measures were computed for each train run to characterize the individual clusters:

- Average, minimum, and maximum delay across stations;
- Range of delays across stations;
- Standard deviation of delays recorded across stations;
- Initial delay, the delay at first station;
- Final delay, the delay at the last station;
- Overall delay change, difference between final and initial delay. Positive values mean the delay has increased from first to last station;
- Maximum delay change across stations.

Cluster characteristics are summarized in Table 4.2-4:

Cluster	N. obs.	Mean average delay [min]	Mean STD of delays [min]	Mean initial delay [min]	Mean final delay [min]	Mean min delay [min]	Mean max delay [min]	Mean delay range [min]	Mean max delay change [min]	Mean overall delay change [min]
1	270	2,26	2,78	-0,95	6,14	-1,06	7,72	8,78	4,88	7,09
2	418	0,55	1,47	-1,05	4,71	-1,24	5,27	6,52	4,69	5,76
3	381	3,09	1,12	1,70	1,80	0,53	4,64	4,11	2,69	0,11
4	159	7,65	1,92	4,59	8,03	3,73	10,21	6,47	6,79	3,44
5	395	1,92	1,14	-0,28	2,23	-0,47	4,10	4,57	2,25	2,51
Total	1623	2,46	1,57	0,35	3,99	-0,12	5,73	5,85	3,87	3,64

Table 4.2-4: Characterization of delay profile clusters, southbound ØK trains Nivå – Copenhagen

4.2.4.2 Comparison with percentile-based approaches on delay profiles

In this section, a comparison is provided between the pooled data and the clusters on the dataset of delay profiles. The same percentile representation of delay profiles is shown, as proposed by Schittenhelm and Richter (2009), Peterson (2012), and van Oort et al. (2015). These authors represented different percentiles. For the sake of clarity, only the 15th, 50th and 85th percentiles and the average are displayed in the following diagrams.

Figure 4.2-3 shows the distribution of delays of the entire dataset of ØK southbound trains. The only pattern visible is a slight increase in delay toward Copenhagen, more evident for the more delayed trains, represented by the 85th percentile. Even though a large portion of punctual trains was discarded from the dataset, the residual distribution of delays remains positively skewed, as shown by the average constantly higher than the median value.

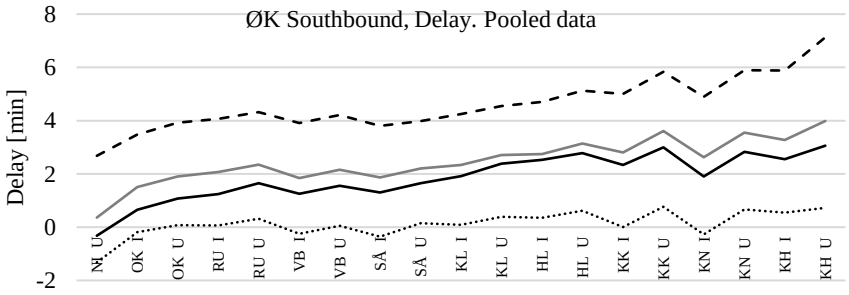

Figure 4.2-3: Delays recorded for ØK southbound trains.15th percentile dotted, median solid black, and 85th percentile dashed. Average solid gray.

The new information revealed by the clustering algorithm is provided in Figure 4.2-4. In this figure, the individual internal distributions of delays are compared to the pooled delay distribution from Figure 4.2-3. Figure 4.2-4 shows, for each cluster, the difference between the cluster statistic at each station and the equivalent pooled statistic from Figure 4.2-3.

In Figure 4.2-4, the 15*th* and 85*th* percentiles and the median line of the internal cluster delay profiles distributions, are compared to the distribution of pooled delay profiles. The clusters where the difference of 85th percentile from the pooled dataset is lower than the difference of the 15th percentile have tighter distributions of delay profiles compared to the pooled dataset, increasing the significance of the identified pattern. The local deviation present in the clusters represents the information hidden in the pooled dataset, which is instead brought to light by the clustering.

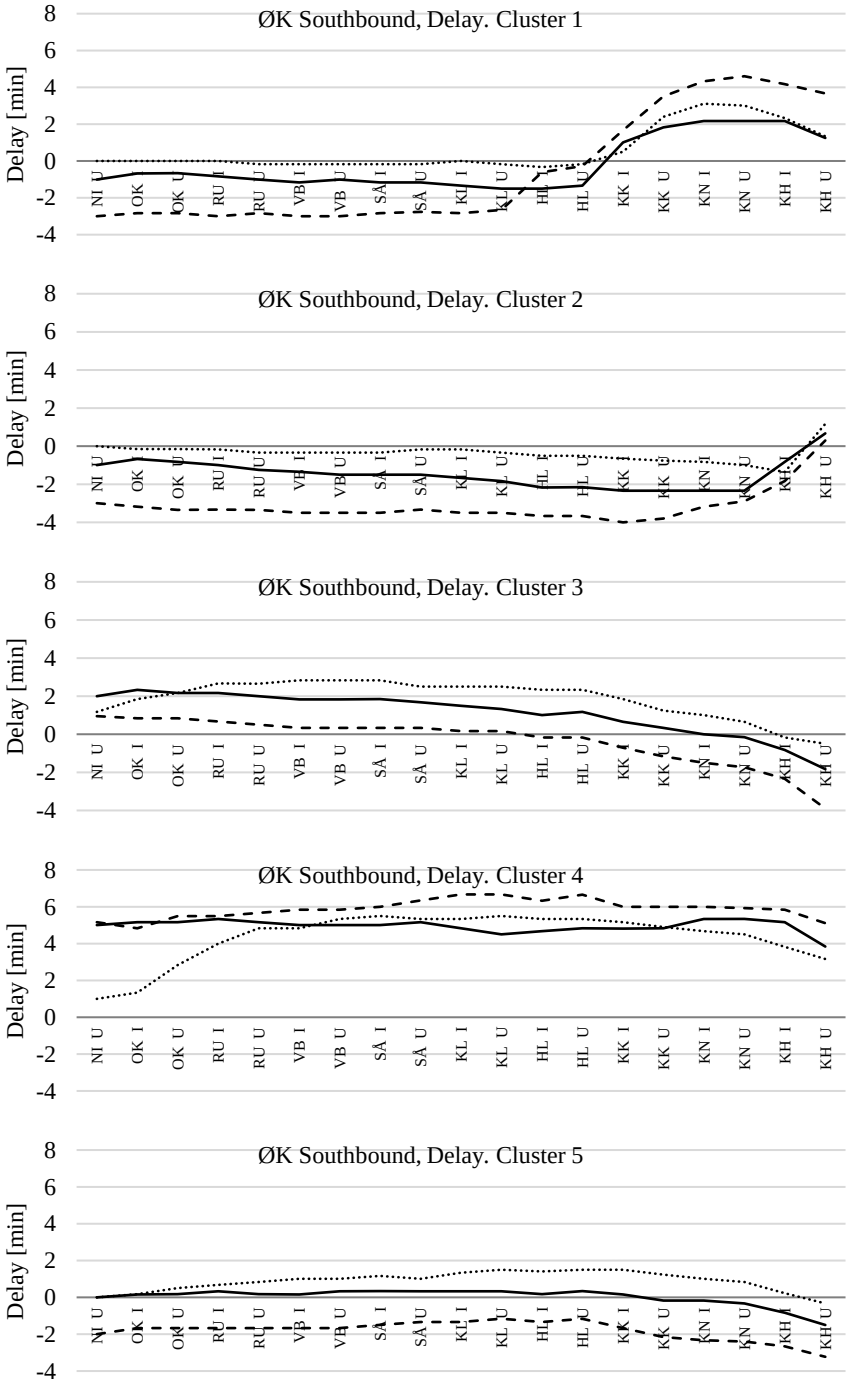

Figure 4.2-4: Differences in the distributions of delays recorded for ØK southbound trains. Each cluster's internal distribution is compared to the pooled distribution. 15th percentile dotted, median solid, and 85th percentile dashed.

4.2.4.3 Comparison with percentile-based approaches on delay change profiles

In this section, a comparison is provided between the pooled data and the clusters on the dataset of delay change profiles. The same representation of delay change profiles based on the median is shown, as proposed by Schittenhelm and Richter (2010; 2009), supplemented with the average, i.e. the cluster centroid.

Figure 4.2-5 shows the delay change profiles of the entire dataset of ØD northbound trains. A generalized positive delay change is visible at the last station. The large changes in delay from location KN I to KK I are linked to the known deviation in the timestamps at Nørreport.

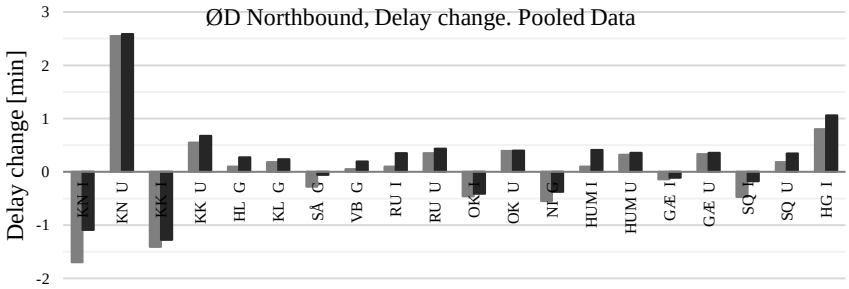

Figure 4.2-5: Delay changes recorded for ØD northbound trains. Median in bright shade, average in dark.

The differences between the pooled median and average delay change profile and the same profiles from individual clusters are represented in Figure 4.2-6. In this case, the information gained by clustering is more evident. All the clusters remain similar to the pooled data at most stations, except few stations, where a large difference is recorded in the delay change.

Every cluster is characterized by at least one larger delay change at one station, which would be hidden in the pooled distribution of delay change profiles. Noticeably, the negative effect of different delay patterns overlapping is evident for *KN I* records. All the clusters deviate negatively from the pooled data by around 0,5 minutes, except for cluster 2, which deviates positively by around 1,5 minutes from the pooled profile. This means that the pooled profile was shifted by one single cluster to a central value, hiding both the frequent delay recovery, and the delay increase specific from cluster 2.

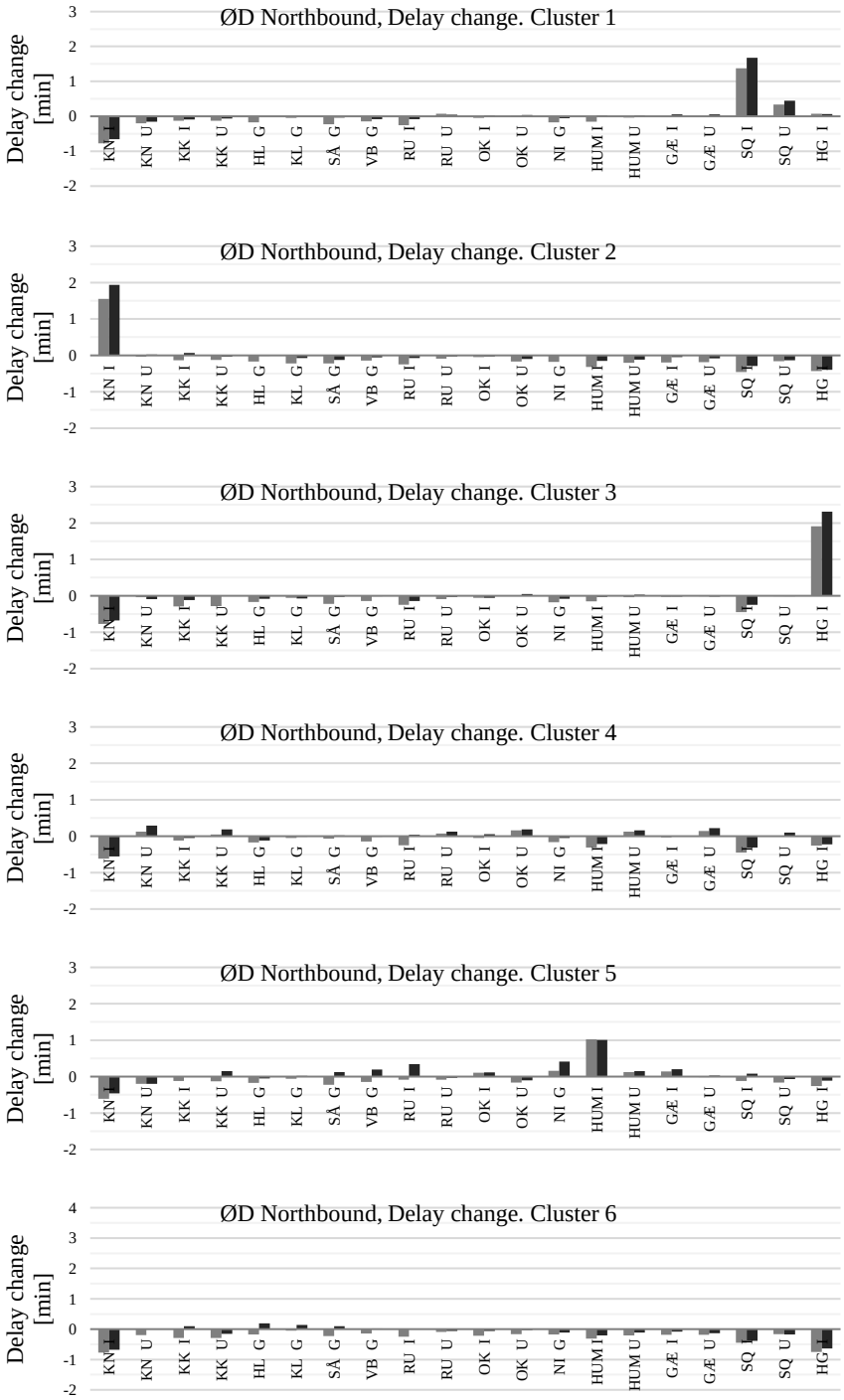

Figure 4.2-6: Delay changes recorded for ØD northbound train, by clusters. Median in bright shade, average in dark.

4.2.4.4 Inference on the clusters

In this section, results from clustering of delay profiles and delay change profiles are investigated to identify relations with cluster characteristics, using heuristic classification. For the sake of conciseness, only cluster centroids are reported in the following figures, and only a sample of the results is reported, which is ØD northbound trains and ØK southbound trains. Figure 4.2-7 shows results from clustering delay change profiles for ØD trains to Helsingør.

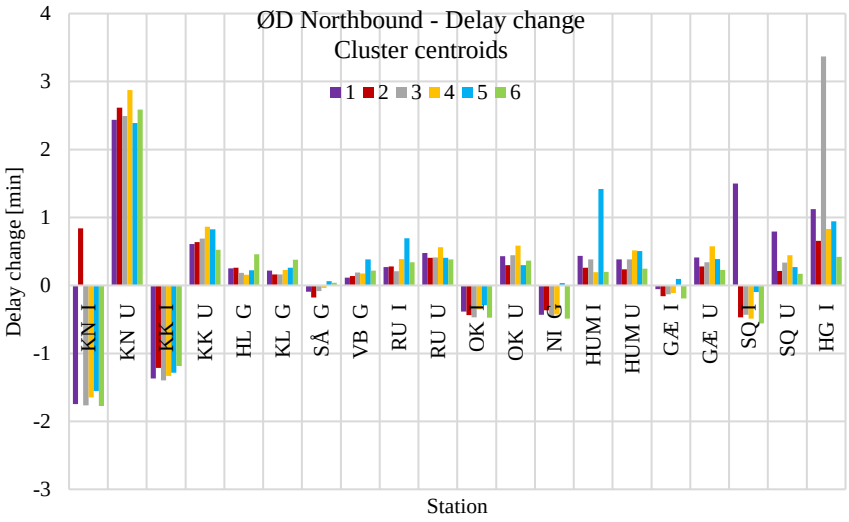

Figure 4.2-7: Cluster centroids for northbound ØD trains, delay change.

Clusters can be interpreted as follow:

1. Cluster 1: regular delay increases at the last three stations, where trains become unpunctual;
2. Cluster 2: delay increase arriving at the first stop, Nørreport;
3. Cluster 3: trains that are considerably delayed arriving at the final destination, Helsingør;
4. Cluster 4: trains without remarkable delay changes: these train tend to keep the same delay throughout the whole journey;
5. Cluster 5: specific delay increases at Humlebæk arrival; trains in this cluster show also smaller recovery at Skodsborg arrival, compared to other clusters;

6. Cluster 6: these trains accumulate delays passing the stations of Hellerup and Klampenborg; on the other side, compared to other clusters, the average delay increase at final destination Helsingør is smaller.

Inference on the cluster population shows that some patterns are specific of selected train services, identified by their train number. Table 4.2-5 shows how every train service ID is spread across clusters. In each column, the shade represents the difference between individual percentages and the cluster share, where the brightest colors are associated with the values furthest from the cluster share. Green is a positive difference, i.e. larger percentages than the cluster share, red is a negative difference, i.e. smaller percentages than the cluster share.

Time band	Departure time from KH	Train number	Cluster 2	6	4	1	3	5
2 – AM Peak	06:18	4413	18%	25%	14%	14%	18%	11%
	06:38	4415	4%	36%	4%	16%	32%	8%
	06:58	4417	26%	33%	7%	19%	4%	11%
	07:18	4419	6%	22%	8%	31%	8%	25%
4 – PM Peak	15:18	4467	21%	25%	17%	4%	21%	13%
	15:38	4469	19%	30%	7%	12%	23%	9%
	15:58	4471	44%	16%	16%	4%	8%	12%
	16:18	4473	5%	15%	28%	18%	20%	15%
	16:38	4475	43%	13%	21%	13%	10%	2%
	16:58	4477	20%	15%	39%	9%	2%	15%
	17:18	4479	16%	32%	15%	12%	9%	16%
	17:38	4481	31%	14%	19%	19%	5%	12%
	17:58	4483	46%	14%	14%	6%	10%	10%
		Cluster share	**24%**	**22%**	**17%**	**14%**	**12%**	**12%**

Table 4.2-5: Northbound ØD trains. Cluster share by train service ID. The color code compares the individual row's distributions among clusters to the overall distribution among clusters reported in the last row. Clusters sorted by size.

Delay change profiles in cluster 1 and 5 represent typical behavior of service 4419, whereas cluster 2 shows considerably more frequent in services 4471, 4473, and 4483. Cluster 3 is more common among services 4415, three times more frequent than the whole population distribution across clusters, and, 4467, 4469, 4473, which double the average frequencies. Cluster 4 is typical for services 4477, and, lastly, Cluster 6 represents

a large share of services 4417 and, again, 4415. Further investigation of other factors may reveal the causes that rule the train services' cluster membership.

The analysis of Table 4.2-5 shows the existence of a relation between train IDs in a specific time band and cluster membership. This is shown in detail in Table 4.2-6, where cluster membership is aggregated in time bands. The same color coding as Table 4.2-5 is applied.

The timetable is divided in time bands according to the overall service frequency on the line, so that time bands 2 and 4 are the AM and PM peak periods, respectively, when 9 trains/h per direction are operated. Time band 1, 3, and 5 are the remaining off-peak periods, when ØD trains are not operated, so only 6 trains/h occupy the line in each direction, allowing for larger headway buffers between trains. At the same time, smaller congestion is expected, in off-peak periods, both on the train traffic and on the number of passengers to board or alight at the stations.

Time	Type	Time band	Cluster					
			2	6	4	1	3	5
6:20 - 8:20	Peak AM	2	13%	28%	9%	21%	15%	15%
15:20 - 18:00	Peak PM	4	27%	20%	19%	12%	11%	11%
		Cluster share	**24%**	**22%**	**17%**	**14%**	**12%**	**12%**

Table 4.2-6: Northbound ØD trains. Cluster share by time band. The color code compares the individual row's distributions among clusters to the overall distribution among clusters reported in the last row. Clusters sorted by size.

In this case, morning peak shows recurrent delay patterns presented by clusters 1 and 6, whereas patterns represented by clusters 2 and 4 are rare in this time band. As opposed, the distribution of trains in the PM peak hour is similar to the overall distribution.

Further inference on the clusters of ØD northbound trains might highlight interferences from other trains. Lokaltog trains run mostly on a network independent from Banedanmark's, and share with ØD and ØP trains the line section between Snekkersten and Helsingør. ØD northbound trains are scheduled at a short headway after Lokaltog trains from Snekkersten to Helsingør. The analysis of timestamps from Lokaltog trains on this section and of the realized headways between Lokaltog and ØD northbound trains might suggest that clusters 1 and 3, which increase the delay near Helsingør, are actually the result of delay propagation from Lokaltog trains to ØD trains.

The clustering results from other service categories, with different stopping patterns, can be related to the time periods of the day. For example, ØK southbound trains are reported in Figure 4.2-8 and Table 4.2-7.

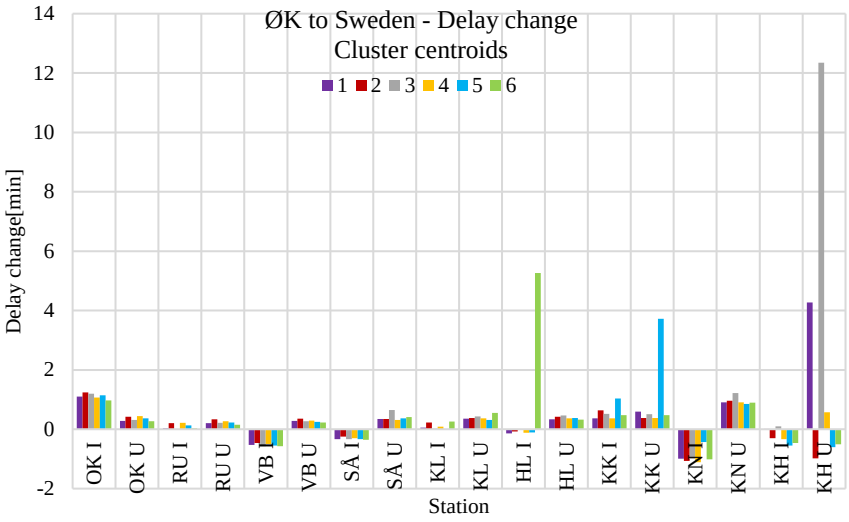

Figure 4.2-8: Cluster centroids for southbound ØK trains, delay change.

Start time	Type	Time band	Cluster 2	4	1	5	3	6
04:30	Off peak	1	35%	30%	15%	12%	3%	5%
06:20	Peak AM	2	49%	26%	10%	11%	1%	3%
08:20	Off peak	3	39%	29%	13%	13%	3%	3%
15:20	Peak PM	4	30%	26%	24%	9%	9%	2%
18:00	Off peak	5	40%	29%	18%	9%	3%	1%
		Cluster share	**40%**	**28%**	**15%**	**12%**	**3%**	**2%**

Table 4.2-7: Southbound ØK trains. Cluster share by time band. The color code compares the individual row's distributions among clusters to the overall distribution among clusters reported in the last row. Clusters sorted by size.

Figure 4.2-8 represents the centroids of resulting clusters in train category ØK Southbound, according to delay change. Besides, the distribution of trains across clusters is summarized in Table 4.2-7, disaggregated by time bands, highlighted in the same color code as Table 4.2-5 and Table 4.2-6. Note that the number of clusters in the delay change profiles of ØK southbound trains is different from the number of clusters in delay profiles. This is not necessarily inconsistent, as the two variables express different aspects of the development of delays. In this case, the cluster share by time band explains the nature of two clusters. In particular, cluster 1 is considerably more frequent in the PM peak hour,

whereas cluster 2 is more typical of trains in the AM peak hour. This result can be reasonably interpreted as delays generated by passenger congestion. In fact, delay increases in the PM peak hour appear at departures from Copenhagen, where a large number of passengers leave towards Sweden. On the contrary, cluster 2 represents delays increases collected across stations towards Copenhagen, and a delay recovery departing from Copenhagen, where fewer passengers are expected to board. The cluster share for clusters 3 to 6 is comparable with the overall distribution across different time bands, so these delay patterns cannot directly be associated with the time of the day. Further research may reveal factors that rule the cluster membership for these clusters.

More disaggregated analysis of cluster composition according to train number, or service ID, is in accordance with aggregated time bands. This is valuable especially for time band 3, which is the most populated time band according to the timetable. Table 4.2-8 shows that, even if the overall distribution of trains in time band 3 across clusters is very similar to the overall distribution, specific train services present different typical delay patterns. In this case, further analysis of train service characteristics should indicate a better disaggregation of train services in a specific time band. The same color code as tables Table 4.2-5, Table 4.2-6 and Table 4.2-7 is applied in Table 4.2-8.

Even though recurrent patterns are also clear in the delay profiles dataset, the results could not be explained by the available variables. Further research might identify relations that guide the clustering of delay profiles on this line, such as realized headways, weather conditions, passenger counts, and recorded delay causes.

Time band	Departure time from NI	Train number	Cluster 2	4	1	5	3	6
1	06:01	1314	38%	20%	23%	13%	3%	5%
	06:21	1316	32%	44%	8%	8%	4%	4%
	06:41	1318	33%	33%	10%	14%	5%	5%
2	07:01	1320	59%	12%	12%	15%	0%	2%
	07:21	1322	50%	29%	13%	4%	0%	4%
	07:41	1324	51%	31%	10%	5%	0%	3%
	08:01	1326	41%	32%	14%	11%	2%	2%
	08:21	1328	53%	18%	0%	22%	2%	5%
	08:41	1330	42%	33%	14%	3%	3%	6%
3	09:01	1332	43%	30%	4%	20%	0%	4%
	09:21	1334	39%	32%	16%	13%	0%	0%
	09:41	1336	41%	24%	15%	17%	0%	2%
	10:01	1338	24%	24%	27%	16%	5%	3%
	10:21	1340	31%	39%	19%	3%	0%	8%
	10:41	1342	57%	14%	11%	14%	0%	5%
	11:01	1344	45%	24%	12%	15%	0%	3%
	11:21	1346	40%	47%	10%	3%	0%	0%
	11:41	1348	53%	35%	7%	5%	0%	0%
	12:01	1350	26%	29%	21%	15%	3%	6%
	12:21	1352	47%	32%	8%	8%	0%	5%
	12:41	1354	38%	16%	28%	16%	0%	3%
	13:01	1356	46%	19%	15%	8%	8%	4%
	13:21	1358	50%	31%	15%	0%	4%	0%
	13:41	1360	59%	24%	7%	3%	7%	0%
	14:01	1362	29%	34%	17%	17%	3%	0%
	14:21	1364	44%	32%	12%	4%	8%	0%
	14:41	1366	37%	37%	10%	15%	2%	0%
	15:01	1368	24%	24%	16%	30%	4%	1%
	15:21	1370	43%	25%	6%	14%	10%	2%
	15:41	1372	35%	41%	5%	11%	5%	3%
	16:01	1374	36%	21%	17%	17%	8%	2%
	16:21	1376	36%	33%	14%	6%	6%	6%
4	16:41	1378	37%	27%	22%	7%	5%	2%
	17:01	1380	19%	29%	29%	13%	6%	3%
	17:21	1382	39%	18%	18%	5%	18%	0%
5	17:41	1384	45%	29%	23%	3%	0%	0%
	18:01	1386	18%	43%	30%	8%	3%	0%
	18:21	1388	47%	28%	16%	9%	0%	0%
	18:41	1390	40%	24%	16%	12%	4%	4%
	19:01	1392	58%	15%	13%	15%	0%	0%
	19:21	1394	34%	31%	17%	10%	7%	0%
	19:41	1396	48%	26%	7%	7%	11%	0%
	20:01	1398	31%	38%	22%	6%	0%	3%
		Cluster share	**40%**	**28%**	**15%**	**12%**	**3%**	**2%**

Table 4.2-8: Southbound ØK trains. Cluster share by service ID. The color code compares the individual row's distributions among clusters to the overall distribution among clusters reported in the last row. Clusters sorted by size.

4.2.5 Discussion

The clustering method proposed in this paper finds its strengths in being automatic, unbiased, flexible, and simple. A comparison to methods presented in the literature is provided in this section. Previous approaches (Andersson et al., 2013b, 2011; Peterson, 2012; Richter, 2010; Schittenhelm and Richter, 2009; van Oort et al., 2015) extracted information from delay profiles mainly through observation, occasionally combined with multiple univariate statistical analyses and observation ranking. In most studies, the complete dataset was plotted in the form of delay profiles, and the identification of frequent patterns among the observations relied on the observer's ability. Visual inspection is typically affected by subjective interpretation, which can differ across analysts, and suffers from the low effectiveness of naked eye to average data represented graphically. In some studies, supporting measures were plotted with the full dataset, such as average profile, median, and selected percentiles to represent the distributions.

The application of these measures as multiple univariate distributions, though, does not catch the interdependencies of delays at different stations and does not provide information about the development of delays along the train journey. The method proposed in this paper allows automatic identification of delay patterns, removing, thus, the influence of subjective interpretation of delay profiles. Furthermore, profile clustering allows the identification of similar delay profiles in the entire pool of records. Note that, even though the clustered delay profiles were plotted in this paper, the observation of the profiles did not play a role in the identification of similarities. This exact process is indeed performed by the clustering algorithm, and the results are then plotted for an easier comprehension of the development of delays in the individual clusters. The metrics provided as 15^{th}, 50^{th}, and 85^{th} percentile would be sufficient to describe the distributions within individual clusters and might be used in replacement of the cluster plots.

Compared to big-data techniques proposed in the literature for other purposes in analysis of transport operation (Goverde and Meng, 2011; Hansen et al., 2010; Kecman and Goverde, 2015, 2012), this method relies on readily available data, and does not need detailed knowledge on the infrastructure and occupation of individual blocking sections. It can, therefore, be scaled to different levels of detail or transferred to other modes of transportation where delay can be measured at fixed points on a given path, such as bus networks or air traffic. It is a very common practice of transport operators to provide live data on delays recorded on their own network, which can be recorded accessing public websites. Furthermore, the partition of operation into recurrent delay patterns allows

inference on individual clusters, which is not possible with association or succession rules (Cule et al., 2011; Wallander and Mäkitalo, 2012; Yabuki et al., 2015). These methods do not provide causality connection, and can only be used to compare scenarios, e.g. before and after delay mitigation countermeasures have been implemented. Results from clustering can be inferred with other mining techniques to identify further connections between specific system factors and delay membership so that the causes of specific delays can be identified, and the effects of corrective actions can be estimated beforehand.

Alongside flexibility, the strength of this method resides in its simplicity. Unsupervised learning methods, such as clustering, aim at the identification of internal structures of the system. Supervised learning methods, in contrast, attempt to predict results, based on assumed connections in the input. For these reasons, neural networks (Chapuis, 2017; Malavasi and Ricci, 2001), Bayesian networks (Kecman et al., 2015), and supporting vector regression methods (Marković et al., 2015) require initial assumptions on the factors that have direct effect on the desired output, which can be cumbersome to identify, and could be hidden. The clustering method proposed here does not require initial assumptions, so any recurrent delay pattern can be identified. In particular, the k-means algorithm was selected, being the most common algorithm for partitional clustering. Even though several clustering methods and algorithms exist in the literature, none of them is clearly preferred from the others (Jain, 2010). It is important to stress the fact that the output of clustering algorithms only suggests hypotheses, and that the interpretation of results plays a more relevant role than seeking the best clustering principle. However, further research might improve the method. For example, a different choice of the clustering statistic between observations might be explored. In addition, the choice of the parameter k might be supported by advanced techniques and metrics. In this paper, k was set through statistical analysis of the associated clusters, but further studies might reveal more efficient methods integrated into the clustering algorithm itself. Lastly, the clustering results might depend on the punctuality threshold selected to filter out punctual trains, if applied.

The relations found in inference from resulting clusters can, eventually, be considered and implemented in the mentioned supervised data mining methods. The use of other sources of information can be further investigated, e.g. the rolling stock equipment deployed, or information on delay causes collected by train dispatchers. The clustering algorithm itself cannot provide information on the causes of delays, but relevant relationships with external variables might be found through the inference on the clusters.

The implementation of information recorded by the dispatchers on primary and secondary delays could support the identification of delay propagation. However, previous studies in Europe highlighted the unreliability of such manually recorded data (Goverde and Meng, 2011; Sørensen et al., 2017). These procedures are different for each infrastructure manager and should comply with different national regulations. This input should be analyzed in detail before being implemented in the inference on clusters. The timestamps might be integrated with data from other railway undertakings so that the realized headways could be investigated and included in the cluster inferences. The effects of delay propagation might be thus investigated, and the dispatching strategies possibly improved. Passenger counts or boarding/alighting timings could also reveal that specific localized delay increases are linked to passenger exchange and might suggest modifications in the scheduled dwelling times. Useful information from the railway undertakings might include differences between planned and realized train compositions or the use of energy saving strategies. Driving support systems are spreading among train operators to reduce energy consumption and thus the operating cost, especially for diesel-powered railways. The effects of such systematic patterns in the driving style are, in any event, expected to emerge in the clustering algorithm, especially with more detailed data in the positioning. Further development of this method might expand its application to other industrial processes or other transportation modes. The service timekeeping could be measured at designated check-points, to build standard delay profiles and delay change profiles.

4.2.6 Conclusions

In this paper, a new method is presented to analyze railway operations, based on big-data techniques. Previous studies highlighted the need for tools to analyze railway operation based on data from automatic data collection sources, and to automatically detect delay patterns (Schittenhelm and Richter, 2009). K-means clustering is here applied to train delay records from automatic train detection systems to identify systematic delays, rearranged in delay profiles and delay-change profiles. This method is automatic, unbiased, flexible, and simple.

Both institutions and industry are showing great interest in big-data applications (D'agostino, 2016). The method described in this paper provides a managerial tool to identify recurrent delay patterns that affect the service reliability. A localized analysis with additional information supports the identification of the causes of individual patterns, so that specific countermeasures can be designed. For example, dispatching strategies might be modified when a structural conflict is detected, the boarding and alighting process might

be improved at stations were delay increases recurrently. If the causes of recurrent delays are identified in frequent conflicts, small modifications to the timetable slack might be a solution to reduce delay propagation.

The effectiveness of this approach is demonstrated in an application on a Danish regional railway line. The application shows that it is possible to identify systematic delays at specific stations in a congested area and to identify different delay patterns. Furthermore, delay patterns can be conveniently associated with specific time periods of the day, showing time dependency, reasonably explained by the prevailing passenger flow direction. Specific delay patterns are demonstrated to be characteristic of individual train service IDs, which could depend on other service characteristics, such as structural conflicts with other trains in specific sections of the line, use of specific rolling stock equipment, or connections to other transport services. The implementation of other sources of information might improve the inference on the clusters, such as weather conditions, passenger counts, information from the dispatchers, or rolling stock characteristics.

Further development of this method might improve the selection of the number of clusters, identify new clustering metrics between observations, or integrate additional sources of information to improve the inference on clusters.

ACKNOWLEDGMENT: This work was funded by a Dean Grant from the Technical University of Denmark (DTU) and by the Danish Innovation Fund through the IPTOP project (Integrated Public Transport Optimisation and Planning).

References

Andersson, E., Peterson, A., Törnquist Krasemann, J., 2013. Introducing a New Quantitative Measure of Railway Timetable Robustness Based on Critical Points, in: Proceedings of 5th International Seminar on Railway Operations Modelling and Analysis (IAROR): RailCopenhagen2013. Copenhagen, pp. 1–19.

Andersson, E., Peterson, A., Törnquist Krasemann, J., 2011. Robustness in Swedish Railway Traffic Timetables, in: Ricci, S., Hansen, I.A., Longo, G.L., Pacciarelli, D., Rodriguez, J., Wendler, E. (Eds.), Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis. Rome, pp. 1–18.

Carey, M., 1999. Ex ante heuristic measures of schedule reliability. Transp. Res. Part B Methodol. 33, 473–494. doi:10.1016/S0191-2615(99)00002-8

Chapuis, X., 2017. Arrival Time Prediction Using Neural Networks, in: Tomii, N., Hansen, I.A., Rodriguez, J., Pellegrini, P., Dauzère-Pérès, S., De Almeida, D. (Eds.), 7th International Conference on Railway Operations Modelling and Analysis. International Association of Railway Operations Research, Lille (France), pp. 1500–1510.

Cule, B., Goethals, B., Tassenoy, S., Verboven, S., 2011. Mining train delays. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 7014 LNCS, 113–124. doi:10.1007/978-3-642-24800-9_13

D'agostino, A., 2016. Big Data in Railways - Common Occurrence Reporting Programme, European Union Agency for Railways. doi:ERA-PRG-004-TD-003

De Mauro, A., Greco, M., Grimaldi, M., 2016. A formal definition of Big Data based on its essential features. Libr. Rev. 65, 122–135. doi:10.1108/LR-06-2015-0061

Gorman, M.F., 2009. Statistical estimation of railroad congestion delay. Transp. Res. Part E Logist. Transp. Rev. 45, 446–456. doi:10.1016/j.tre.2008.08.004

Goverde, R.M.P., Hooghiemstra, G., Lopuhaä, H.P., 2001. Statistical Analysis of Train Traffic: The Eindhoven Case, TRAIL studies in transportation science series. IOS Press, Incorporated.

Goverde, R.M.P., Meng, L., 2011. Advanced monitoring and management information of railway operations. J. Rail Transp. Plan. Manag. 1, 69–79. doi:10.1016/j.jrtpm.2012.05.001

Hansen, I.A., Goverde, R.M.P., Van Der Meer, D.J., 2010. Online train delay recognition and running time prediction, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. IEEE, pp. 1783–1788. doi:10.1109/ITSC.2010.5625081

Hansen, J., 2015. Future Railway Development and Performance, in: The Danish Rail Conference. The Danish Rail Sector Association, Copenhagen.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed, Elements. Springer. doi:10.1007/b94608

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31, 651–666. doi:10.1016/j.patrec.2009.09.011

Kashima, H., Hu, J., Ray, B., Singh, M., 2008. K-means clustering of proportional data using L1 distance. IEEE, Tampa, FL, USA, pp. 1–4. doi:10.1109/ICPR.2008.4760982

Kecman, P., Corman, F., Peterson, A., Joborn, M., 2015. Stochastic prediction of train delays in real-time using Bayesian networks, in: Proceedings of the 13th Conference on Advanced Systems in Public Transport (CASPT) 2015. Erasmus University, Rotterdam (Netherlands), p. 18.

Kecman, P., Goverde, R.M.P., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transp. 7, 295–319. doi:10.1007/s12469-015-0106-7

Kecman, P., Goverde, R.M.P., 2012. Process mining of train describer event data and automatic conflict identification, in: Computers in Railways XIII. WIT Transactions on the Built Environment, New Forest, UK, pp. 227–238. doi:10.2495/CR120201

Lindfeldt, A., 2010. A study of the performance and utilization of the Swedish railway network, in: Lakušić, S. (Ed.), First International Conference on Road and Rail Infrastructure - CETRA2010. University of Zagreb, Opatija, Croatia.

Lindfeldt, A., Sipilä, H., 2014. Simulation of freight train operations with departures ahead of schedule, in: Sipilä, H. (Ed.), WIT Transactions on the Built Environment. WIT Press, pp. 239–249. doi:10.2495/CR140191

Lindfeldt, O., 2010. Evaluation of punctuality on a heavily utilised railway line with mixed traffic, in: WIT Transactions on the Built Environment. WIT Press, pp. 545–553. doi:10.2495/CR080531

Malavasi, G., Ricci, S., 2001. Simulation of stochastic elements in railway systems using self-learning processes. Eur. J. Oper. Res. 131, 262–272. doi:10.1016/S0377-2217(00)00126-0

Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P., 2015. Analyzing passenger train arrival delays with support vector regression. Transp. Res. Part C Emerg. Technol. 56, 251–262. doi:10.1016/j.trc.2015.04.004

Olsson, N.O.E., Haugland, H., 2004. Influencing factors on train punctuality—results from some Norwegian studies. Transp. Policy 11, 387–397. doi:10.1016/j.tranpol.2004.07.001

Peterson, A., 2012. Towards a robust traffic timetable for the Swedish Southern Mainline, in: Computers in Railways XIII. WIT Transactions on The Built Environment, New Forest, UK, pp. 473–484. doi:10.2495/CR120401

Richter, T., 2012. Data aggregation for detailed analysis of train delays, in: Brebbia, C.A., Tomii, N., Mera, J.M., Ning, B., Tzieropoulos, P. (Eds.), WIT Transactions on the Built Environment. WITPress, pp. 239–250. doi:10.2495/CR120211

Richter, T., 2010. Systematic analyses of train run deviations from the timetable, in: Computer in Railways XII. Wit Trans B, pp. 651–662. doi:10.2495/CR100601

Richter, T., 2008. En bedre jernbane gennem højere datakvalitet, in: Harry Lahrmann (Ed.), Trafikdage På Aalborgs Universitet. Traffic Research Group, Aalborg University, Aalborg.

Richter, T., Landex, A., Andersen, J.L.E., 2013. Precise and accurate train run data: Approximation of actual arrival and departure times, in: WCRR (World Congress Railway Research). International Association of Railways, Sydney (Australia).

Schittenhelm, B., Richter, T., 2009. Railway Timetabling Based on Systematic Follow-up on Realized Railway Operations, in: Harry Lahrmann (Ed.), Proceedings from the Annual Transport Conference at Aalborg University. Traffic Research Group, Aalborg University, Aalborg.

Schittenhelm, B.H., 2011. Planning With Timetable Supplements in Railway Timetables, in: Annual Transport Conference at Aalborg University. trafikdage, Aalborg, DK.

Shih, M.-C., Dick, C.T., Sogin, S.L., Barkan, C.P.L., 2014. Comparison of Capacity Expansion Strategies for Single-Track Railway Lines with Sparse Sidings. Transp. Res. Rec. J. Transp. Res. Board 2448, 53–61. doi:10.3141/2448-07

Sipilä, H., 2010. Simulation of modified timetables for high speed trains Stockholm – Göteborg, in: Lakušić, S. (Ed.), First International Conference on Road and Rail Infrastructure - CETRA2010. University of Zagreb, Opatija, Croatia, p. 1.

Sørensen, A.Ø., Landmark, A.D., Olsson, N.O.E., Seim, A.A., 2017. Method of analysis for delay propagation in a single-track network. J. Rail Transp. Plan. Manag. 7, 77–97. doi:10.1016/j.jrtpm.2017.04.001

van Oort, N., Sparing, D., Brands, T., Goverde, R.M.P., 2015. Data driven improvements in public transport: the Dutch example. Public Transp. 7, 369–389. doi:10.1007/s12469-015-0114-7

Wallander, J., Mäkitalo, M., 2012. Data mining in rail transport delay chain analysis. Int. J. Shipp. Transp. Logist. 4, 269. doi:10.1504/IJSTL.2012.047492

Wu, J., 2012. Advances in K-means Clustering, Springer Theses. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1017/CBO9781107415324.004

Yabuki, H., Ageishi, T., Tomii, N., 2015. Mining the Cause of Delays in Urban Railways based on Association Rules, in: CASPT2015. Rotterdam, pp. 1–16.

The reliability of railway transport is one of the key factors to ensure its attractiveness.

This thesis investigates the phenomena related to delays in railways, from both theoretical and empirical perspectives. Firstly, the study evaluates a set of ex-ante measures to estimate the reliability of a timetable, divided into analytical and simulation-based measures. Secondly, an analytical delay propagation model is developed to assess the service reliability, combining the velocity of analytical methods and the accuracy of simulation-based methods. Finally, empirical studies of the realized operation on the Danish railway network are introduced to estimate the input parameters for the analytical delay propagation model. In addition, the analysis of historical data reveals recurrent delay patterns to focus the mitigation actions for improving the service reliability.

DTU Management Engineering
Department of Management Engineering
Technical University of Denmark

Produktionstorvet
Building 424
DK-2800 Kongens Lyngby
Denmark
Tel. +45 45 25 48 00
Fax +45 45 93 34 35

www.man.dtu.dk